

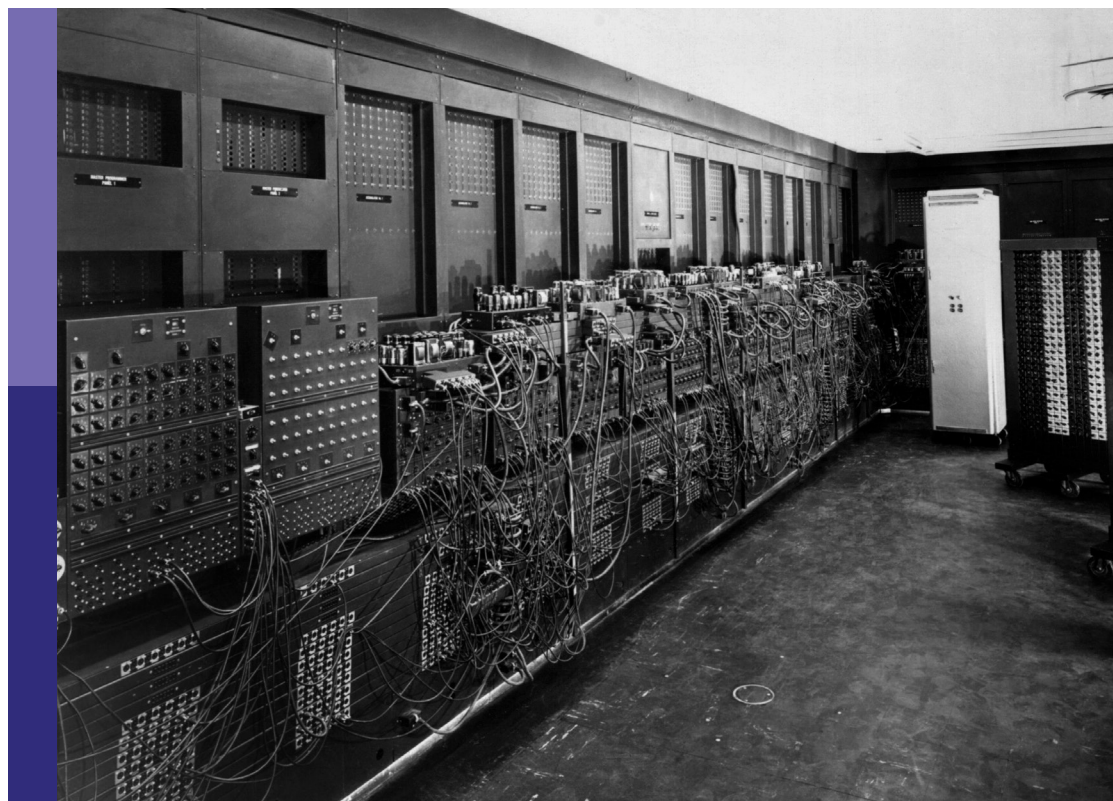
# Drawing multimodality's bigger picture: Metalanguages and corpora for multimodal analyses

**Edited by**

Janina Wildfeuer and Claudia Lehmann

**Published in**

Frontiers in Communication



## FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714  
ISBN 978-2-8325-5196-7  
DOI 10.3389/978-2-8325-5196-7

## About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: [frontiersin.org/about/contact](https://frontiersin.org/about/contact)

# Drawing multimodality's bigger picture: Metalanguages and corpora for multimodal analyses

## Topic editors

Janina Wildfeuer — University of Groningen, Netherlands

Claudia Lehmann — University of Potsdam, Germany

## Citation

Wildfeuer, J., Lehmann, C., eds. (2024). *Drawing multimodality's bigger picture: Metalanguages and corpora for multimodal analyses*. Lausanne: Frontiers Media SA.  
doi: 10.3389/978-2-8325-5196-7

# Table of contents

05	<b>Editorial: Drawing multimodality's bigger picture: metalanguages and corpora for multimodal analyses - in lieu of a Festschrift for John A. Bateman</b> Janina Wildfeuer and Claudia Lehmann
08	<b>Intradisciplinarity: can one theory do it all?</b> J. R. Martin
22	<b>Rethinking multimodal corpora from the perspective of Peircean semiotics</b> Tuomo Hiippala
37	<b>What makes a multimodal construction? Evidence for a prosodic mode in spoken English</b> Claudia Lehmann
52	<b>Fresh perspectives on multimodal argument reconstruction</b> Hartmut Stöckl
58	<b>Approaching tourism communication with empirical multimodality: exploratory analysis of Instagram and website photography through data-driven labeling</b> Elena Mattei
67	<b>Refining concepts for empirical multimodal research: defining <i>semiotic modes</i> and <i>semiotic resources</i></b> Jacopo Castaldi
79	<b>The materiality key: how work on empirical data can improve analytical models and theoretical frameworks for multimodal discourse analysis</b> Arianna Maiorani
92	<b>Multimodal cohesion and viewers' comprehension of scene transitions in film: an empirical investigation</b> Dayana Markhabayeva and Chiao-I Tseng
111	<b>All eyes on the signal? - Mapping cohesive discourse structures with eye-tracking data of explanation videos</b> Leandra Thiele, Florian Schmidt-Borcherding and John A. Bateman
134	<b>World futures through RT's eyes: multimodal dataset and interdisciplinary methodology</b> Anna Wilson, Irina Pavlova, Elinor Payne, Ilya Burenko and Peter Uhrig
155	<b>How films convey meaning through alternating structures (with an illustrative analysis of <i>The Sunbeam</i>)</b> Karl-Heinrich Schmidt
171	<b>Diachronic multimodality research – a mini-review</b> Jana Pflaeging



- 178 **The cognitive roots of multimodal symbolic forms with an analysis of multimodality in movies**  
Wolfgang Wildgen
- 187 **SFDRS as a metalanguage for ‘foodscaping’: adding a formal dimension to an interdisciplinary, multimodal approach to food**  
Loli Kim and Niamh Calway



## OPEN ACCESS

EDITED AND REVIEWED BY  
Hartmut Stöckl,  
University of Salzburg, Austria

\*CORRESPONDENCE  
Janina Wildfeuer  
✉ j.wildfeuer@rug.nl

RECEIVED 22 May 2024  
ACCEPTED 19 June 2024  
PUBLISHED 10 July 2024

## CITATION

Wildfeuer J and Lehmann C (2024) Editorial:  
Drawing multimodality's bigger picture:  
metalanguages and corpora for multimodal  
analyses - in lieu of a Festschrift for John A.  
Bateman. *Front. Commun.* 9:1436821.  
doi: 10.3389/fcomm.2024.1436821

## COPYRIGHT

© 2024 Wildfeuer and Lehmann. This is an  
open-access article distributed under the  
terms of the [Creative Commons Attribution  
License \(CC BY\)](#). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic practice.  
No use, distribution or reproduction is  
permitted which does not comply with these  
terms.

# Editorial: Drawing multimodality's bigger picture: metalanguages and corpora for multimodal analyses - in lieu of a Festschrift for John A. Bateman

Janina Wildfeuer<sup>1\*</sup> and Claudia Lehmann<sup>2</sup>

<sup>1</sup>Center for Language and Cognition, Faculty of Arts, University of Groningen, Groningen, Netherlands,  
<sup>2</sup>Department of English and American Studies, University of Potsdam, Potsdam, Germany

## KEYWORDS

multimodality research, meta-methodology, empirical research, corpus analysis, interdisciplinarity

## Editorial on the Research Topic

[Drawing multimodality's bigger picture: metalanguages and corpora for multimodal analyses - in lieu of a Festschrift for John A. Bateman](#)

## 1 Drawing multimodality's bigger picture

The present Research Topic is dedicated to the work and achievements of one of the key figures in multimodality research, Professor John A. Bateman. John officially retired from his professorship in September 2023, after nearly 25 years of service at the University of Bremen, Germany, and more than 35 years of academic work at several universities and research centers around the world. Those who know John are aware that this retirement is desirably only a formality and that he will hopefully remain engaged in academic work for many years to come.

In good old German academic tradition, professorial retirements are often accompanied by a so-called Festschrift, a “celebratory writing,” i.e. a book honoring the academic and their work, with contributions from colleagues, friends, and PhD students. However, such a Festschrift is something that John himself did not allow to happen due to “too much cult of the individual” (quote from a personal email conversation in 2023). Normally, as long-time colleagues and mentees, we follow John's judgment and accept his decision, but through his training, we also learned to scrutinize and challenge some of his statements, only for his own good, of course.

So here we are with this Research Topic, which is indeed a collection of articles by John's colleagues, research associates, friends, and PhD students, in a journal that is edited by him and that is dedicated to one of his main research fields: multimodal communication. The majority of the articles in this Research Topic are papers from two conference panels we organized for the 11th International Conference on Multimodality (ICOM-11) in London in September 2023; some others are contributions from even more colleagues and friends from all over the world. We thank everyone cordially for their input, support, interest, and contribution to this Research Topic—and we thank John for making it possible. After all, it

was John who officially approved our plans to organize the conference panels and publish this Research Topic as something “almost reasonable” (another quote from a personal email conversation in 2023). Admittedly and naturally, John did not have much say in the matter after his approval, nor was he involved in the review process. Of course, his spirit is present (or alive) in every contribution and the research behind it—and this is exactly what we were aiming for: we wanted to bring together scholars from a variety of disciplines interested in multimodality research to review, explore, and advance the contributions that John has made both to theory- and method-building and to the advancement of multimodal empirical and corpus analyses.

## 2 Metalanguages and corpora for multimodal analysis

Our main starting points for this Research Topic were twofold, building on discussion points recently raised by John himself: On the one hand, after 30 years of development, mainly in the humanities, and after having been evaluated in many different ways (see for an overview Wildfeuer et al., 2019), multimodality should no longer be seen as a research field or discipline, but rather as a “stage of development within a field,” a stage that every discipline goes through (Bateman, 2022a, p. 49). This means that many different fields and disciplines (not only in the humanities) have already entered, are currently entering, or will soon enter their own multimodal phase with a renewed interest in multimodal phenomena. With this comes a particular commitment to theory and method development, with each discipline or field bringing in its own principles and tools. This leads not only to an immense breadth of potential objects of analysis and points of discussion, but also and more importantly to the need to bridge differences and incompatibilities in favor of what John calls a “meta-methodology”:

“We need to find ways of ‘combining’ insights from the variously imported theoretical and methodological backgrounds brought along by previous non-multimodal stages of any contributing disciplines.” (Bateman, 2022a, p. 49)

On the other hand, this search for a meta-methodology to guide multimodal analysis has recently been driven by more empirical approaches and the development and use of larger multimodal corpora, which also require theoretical and methodological refinement.

“We need to develop ways of strengthening claims with robustly applicable methods which nevertheless remain firmly anchored theoretically.” (Bateman, 2022b, p. 64)[SIC]

Making available these large-scale corpora and providing broader and more complex empirical and experimental setups aim to reconceptualize the practice of multimodal analysis and fully implement the “move from theory to data” (see Pflaeging et al., 2021). Following Bateman (2022a), for a productive treatment of these issues, disciplinary triangulation and the development of a “common language” or metalanguage (Maton and Chen, 2016) for an “integrationist interdisciplinarity” (Van Leeuwen, 2005) are the

greatest challenges in contemporary multimodality research. It is precisely these challenges that we productively defined as the main aims of this Research Topic and as “a multimodal task from the ground up” (Bateman, 2022b, p. 64). We explicitly called for works that critically addressed John’s theoretical and methodological advancements, that tested and reviewed the many approaches that he has developed for the analysis of multimodal artifacts, and that expanded on or even rejected some of the ideas and insights provided in his work.

True to John’s research, the resulting contributions show theoretical and methodological concerns on the one hand, and data-driven analyses and approaches to a variety of multimodal artifacts on the other. Similar to the breadth and depth of his own work in more than 350 publications since 1983, the contributions to the present Research Topic are diversely rich and broad, ranging from brief research reports to a mini review to expanded research articles, all of which make a significant contribution to the field of multimodality research. Several articles challenge the theoretical and methodological concepts that were originally discussed and/or further developed by John, such as the notion of discourse semantics and a multimodal metalanguage (Martin), the concept of semiotic mode (Castaldi), the use of Segmented Discourse Representation Theory for multimodal artifacts (Kim and Calway), or the idea of a comprehensive semiotics for multimodal (corpus) analysis (Wildgen; Hiippala). Some papers show the breadth and reach of these theoretical and methodological concepts to provide an application-oriented approach to specific sub-disciplines of multimodality research, including diachronic multimodality studies (Pflaeging), multimodal argumentation studies (Stöckl), or multimodal corpus analysis (Hiippala). Several other articles provide results and evidence from empirical multimodality research with annotation systems and/or larger corpora (Maiorani; Hiippala; Thiele et al.), computational and (semi-)automatic tools (Wilson et al.; Mattei), or experimental studies such as eyetracking, surveys and interviews, or motion detection (Thiele et al.; Markhabayeva and Tseng; Lehmann; Maiorani). Together, these contributions provide insights into a wide range of communicative situations and media, including face-to-face interactions (Lehmann; Wilson et al.), foodscaping (Kim and Calway), film and audiovisual media (Wildgen; Thiele et al.; Markhabayeva and Tseng; Schmidt), websites and social media (Mattei), dance (Maiorani), and diagrams (Hiippala).

John can and should be present in all contributions—certainly not as a dominant sovereign (of which he was initially afraid; cf. “the cult of the individual” in Durkheim, 1964). Instead, we believe that each contribution developed its own voice and standpoint as part of the bigger picture of multimodality research. This voice may have been trained, educated, and influenced by John, through his writings, his comments and reviews, or his famous discussion practice, but it is certainly also presented with a particular independent stance, be it critical or affirmative, bringing out new and sometimes challenging ideas, reasonably.

Following the suggestion by Hiippala (2024), we label this Research Topic a “not-a-Festschrift Research Topic,” because it is, indeed, not simply a way of honoring John’s scholarly achievements in a retrospective. Rather, this Research Topic intends to foster theory- and method-building in multimodality research with a prospective, future-oriented, outlook. Very much in the spirit

of John's work as a mentor, supervisor, colleague and friend, we see the papers as examples of intellectual positions that can and should be discussed and challenged. We also see them as calls for future work, for the advancement of the field of multimodality research, something that John has always striven for with admirable curiosity, open-mindedness, and exceptional innovation and commitment.

## Author contributions

JW: Conceptualization, Project administration, Resources, Writing – original draft, Writing – review & editing. CL: Conceptualization, Project administration, Resources, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Acknowledgments

We thank Tamara Drummond for the initial support and setup of this Research Topic and the accompanying conference panels.

## References

- Bateman, J. A. (2022a). Multimodality, where next? – some meta-methodological considerations. *Multimodality Soc.* 2, 41–63. doi: 10.1177/26349795211073043
- Bateman, J. A. (2022b). Growing theory for practice: empirical multimodality beyond the case study. *Multimodal Commun.* 11, 63–74. doi: 10.1515/mc-2021-0006
- Durkheim, E. (1964). *The Division of Labor in Society* (Transl. by G. Simpson). New York, NY: The Free Press.
- Hiippala, T. (2024). *I have a weird new article out in @FrontComm that discusses multimodal corpora from the perspective of Peircean semiotics. It's also a part of John Bateman's not-a-Festschrift special issue edited by @neous et al. #multimodality #digitalhumanities*. Twitter/X. Available online at: [https://twitter.com/tuomo\\_h/status/1756999198292017546](https://twitter.com/tuomo_h/status/1756999198292017546) (accessed June 24, 2024).
- Maton, K., and Chen, R. T.-H. (2016). "LCT in qualitative research: creating a translation device for studying constructivist pedagogy," in *Knowledge-Building: Educational Studies in Legitimation Code Theory*, eds. K. Maton, S. Hood, and S. Shay (London: Routledge), 27–48.
- Pflaeging, J., Bateman, J. A., and Wildfeuer, J. (2021). "Empirical multimodality research: the state of play," in *Empirical Multimodality Research: Methods, Evaluations, Implications*, eds. J. Pflaeging, J. Wildfeuer, and J. A. Bateman (Berlin: De Gruyter), 1–32. doi: 10.1515/9783110725001-001
- Van Leeuwen, T. (2005). "Three models of interdisciplinarity," in *A New Agenda in (critical) Discourse Analysis: Theory, Methodology and Interdisciplinarity*, eds. R. Wodak, and P. Chilton (Amsterdam: John Benjamins), 3–18. doi: 10.1075/dapsac.13.04lee
- Wildfeuer, J., Pflaeging, J., Bateman, J. A., Seizov, O., and Tseng, C. (2019). "Multimodality. disciplinary thoughts and the challenge of diversity – introduction," in *Multimodality. Disciplinary Thoughts and the Challenge of Diversity*, eds. J. Wildfeuer, J. Pflaeging, J. A. Bateman, O. Seizov, and C. Tseng (Berlin: de Gruyter), 3–38. doi: 10.1515/9783110608694

We also thank all contributors for their work in and for the Research Topic. A particular thank you to Damiano La Manna at Frontiers for his continuous support throughout the process of creating, editing, and finalizing this Research Topic.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



## OPEN ACCESS

## EDITED BY

Janina Wildfeuer,  
University of Groningen, Netherlands

## REVIEWED BY

Pauline Jones,  
University of Wollongong, Australia  
Fei Victor Lim,  
Nanyang Technological University, Singapore

## \*CORRESPONDENCE

J. R. Martin  
✉ james.martin@sydney.edu.au

RECEIVED 09 October 2023

ACCEPTED 07 December 2023

PUBLISHED 05 January 2024

## CITATION

Martin JR (2024) Intradisciplinarity: can one theory do it all? *Front. Commun.* 8:1310001. doi: 10.3389/fcomm.2023.1310001

## COPYRIGHT

© 2024 Martin. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Intradisciplinarity: can one theory do it all?

J. R. Martin\*

Department of Linguistics, University of Sydney, Sydney, NSW, Australia

This position paper draws on Bernstein and Maton's sociology of knowledge to explore Systemic Functional Linguistics (SFL) and Systemic Functional Semiotics (SFS), alongside their relation to Bateman's vision for empirical multimodality research. The paper suggests that SFL/SFS's internal grammar is by and large compatible with Bateman's vision, even if its external grammar falls short insofar as extant descriptions of one semiotic system or another are concerned. The paper closes with the suggestion that SFS and Bateman's multimodality can learn most from one another in research projects embracing a dialectic of theory, description, and ideologically committed practice.

## KEYWORDS

Systemic Functional Linguistics (SFL), Systemic Functional Semiotics (SFS), legitimization code theory (LCT), multimodality, transdisciplinarity

## 1 Disciplinarity

In recent papers, [Bateman \(2020a,b\)](#); [Bateman \(2021, 2022a,b\)](#) explores his vision for multimodality as an empirical discipline. In doing so, he draws on sociological studies of knowledge structure, including the work by [Bernstein \(2000\)](#), [Maton \(2011, 2014, 2016\)](#), [Maton and Chen \(2016\)](#), [Maton and Howard \(2016\)](#), and [Maton et al. \(2016\)](#). As part of this projection, he warns against falling foul of "various flavors and variations of Saussure's well-known proposal of language (or any other system) as a 'master template' for semiotics as such" ([Bateman, 2022a](#), p. 47) and what he calls "linguistic imperialism" ([Bateman, 2022b](#), p. 63). In addition, he notes that "predatory" interdisciplinarity "will be rejected from the start" ([Bateman, 2021](#), p. 308).

Read in tandem with Kress's many declarations of a new age of meaning making called "Multimodality" (e.g., [Kress, 2003, 2010, 2015](#)), superseding language and the discipline of linguistics, serious questions have to be raised about the work on multimodality informed by a theory of language such as Systemic Functional Linguistics (SFL)—work evolving into something we might call Systemic Functional Semiotics (SFS) via publications such as Kress and van Leeuwen's *Reading Images* ([Kress and van Leeuwen, 1990](#) and subsequent editions), [Caple \(2013\) Photojournalism](#), [Doran \(2018\) The Discourse of Physics](#), [Painter et al.'s \(2013\) Reading Visual Narratives](#), [He \(2021\) "Toward a stratified metafunctional model of animation,"](#) [Ngo et al. \(2022\) Modeling Paralanguage using Systemic Functional Semiotics](#), [Martin and Unsworth \(2024\) Reading Images for Knowledge Building](#), [Zappavigna and Logi \(2024\) Emoji and Social Media Paralanguage](#), and [Yu \(forthcoming\) Multimodal Knowledge Building in Secondary School Chemistry Textbooks](#).

Accordingly, in this paper, I will draw on the sociological studies referred to above to explore the nature of SFL and SFS as knowledge structures, compare them with the model of empirical multimodality envisioned by Bateman, and make

some suggestions about how his ambitions for the field might be most effectively accommodated. I write as an SFL linguist (discourse analyst in particular), who has been drawn into work on multimodality by research students and colleagues over the past two and a half decades. As such, given the misgivings about the contribution of linguistics noted above, I should perhaps request readers' indulgence—as I suggest that an SFL/SFS perspective need not be read as the foul and predatory one that some of the more logophobic multimodalists apparently fear.

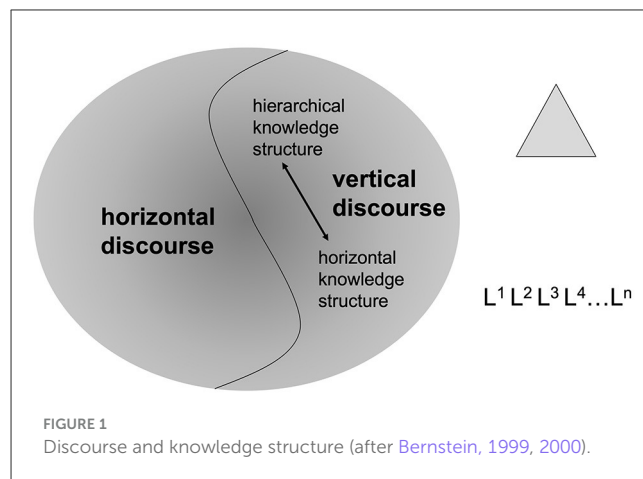
In discussions of this kind, it is important to distinguish multimodality as a field of research and multimodality as its object of study. Multimodalists (like psychologists) unfortunately tend to use the same term for both phenomena (cf., language and linguistics for linguists). Where confusion might arise, I will refer to the field of research as Multimodal Studies below.

## 2 Knowledge structure

By way of framing the discussion, let us begin with Bernstein's (1996, p. 23) distinction between singulars and regions. For Bernstein, a singular is “a discourse which has appropriated a space to give itself a unique name,” for example, “physics, chemistry, sociology, psychology” and which “created the field of the production of knowledge.” These he contrasts with regions, “a recontextualising of singulars,” for example, “medicine, architecture, engineering, information science,” noting that “any regionalisation of knowledge implies a recontextualising principle: which singulars are to be selected, what knowledge within the singular is to be introduced and related.” Importantly, he goes on to comment that “regions are the interface between the field of the production of knowledge and any field of practice.” Had Bernstein's vision extended into the 21st century, he might well have added multimodality as an emerging region to his list, with media and communication as its field of practice.

Seen in these terms, SFL is a canonical singular (Martin, 2014, 2016) and contrasts with its regionalisation in the Sydney School's well-known genre-based literacy programmes (Rose and Martin, 2012)—which tend to draw on a range of relevant singulars (including, for example, Bernstein and Maton's sociology of knowledge, neo/Vygotskian social psychology, and strands of critical discourse analysis). One possible reading of Bateman's vision would entail, via design and/or evolution, the transformation of Multimodal Studies into a singular—with its own distinctive knowledge structure deploying an empirical methodology grounding theory and description.

Bernstein's perspective is further elaborated in the distinction he draws between horizontal and vertical discourse (an opposition between what he earlier referred to as common and uncommon sense). A horizontal discourse involves “a set of strategies which are local, segmentally organized, context specific and dependent, for maximizing encounters with persons and habitats.... This form has a group of well-known features: it is likely to be oral, local, context dependent and specific, tacit, multi-layered and contradictory across but not within contexts” (Bernstein, 2000, p. 157). A vertical discourse on the other hand “takes the form of a coherent, explicit and systematically principled structure, hierarchically organized as in the sciences, or it takes the form of a series of specialized



languages with specialized modes of interrogation and specialized criteria for the production and circulation of texts as in the social sciences and humanities” (Bernstein, 2000, p. 157).

In addition, two forms of vertical discourse are distinguished—hierarchical knowledge structures vs. horizontal ones. A hierarchical knowledge structure is “a coherent, explicit and systematically principled structure, hierarchically organized” which “attempts to create very general propositions and theories, which integrate knowledge at lower levels, and in this way shows underlying uniformities across an expanding range of apparently different phenomena” (Bernstein, 1999, p. 161–162)—e.g., physics, chemistry, or biology. A horizontal knowledge structure, on the other hand, is defined as “a series of specialized languages with specialized modes of interrogation and criteria for the construction and circulation of texts” (Bernstein, 1999, p. 162)—e.g., linguistic theories which position themselves as functional, arguably West Coast Functionalism, Lexical Functional Grammar, Functional Grammar, Discourse Functional Grammar, Role and Reference Grammar or Systemic Functional Linguistics. Bernstein uses a triangle to symbolize hierarchical knowledge structures since they attempt to create ever more general propositions which account for an expanding range of phenomena (e.g., Newtonian physics, superseded by Einstein's relativity, and superseded by string theory). Horizontal knowledge structures, on the other hand, are visualized by a succession of “Ls” since what counts as development is the introduction of a new perspective, typically by junior speakers who challenge the power and legitimacy of more senior ones (e.g., Marxist history, feminist history, and post-colonial history). A synoptic overview of these distinctions is offered in Figure 1.

As exemplified above, in Bernstein's terms, SFL is a canonical member of a horizontal knowledge structure comprising many different theories. Bateman's vision for Multimodal Studies is perhaps a more ambitious one, leaning toward the design and evolution of a hierarchical knowledge structure. This is a trajectory that linguistic theories have embraced, without success, since the modern discipline was founded by Saussure (1916/1959).

Wignell (2007a,b) examines the history of social science, focussing on the emergence of economics, political economy, and sociology as “a hybrid of the language of the physical



sciences and the language of the humanities” (Wignell, 2007a, p. 202)—suggesting that the stronger the boundaries around one of these disciplines, the more it will evolve the characteristics of a hierarchical knowledge structure. In his 2004 conference presentation of Wignell (2007a), he in fact refers to social science knowledge structures as “warring triangles,” since they, in general, aspire to be recognized as hierarchical knowledge structures (viz., linguists’ claims for their discipline as the “science of language”). What happens in practice, however, is that one or another linguistic theory gains institutional rather than intellectual control of the discipline, for a specific period of time, in a specific place (e.g., Chomskyan linguistics’ supremacist control of American linguistics and its intellectual dominions in the 1960s, waning not long thereafter). Seen in these terms, Bateman’s vision involves strengthening boundaries around what counts as empirical Multimodal Studies, thereby fostering its development as a hierarchical knowledge structure—occluding more “weakly bounded” competing triangles as it does so and enjoying globalized longevity.

Bernstein (2000, p. 132–134) probes more deeply into the characteristics of hierarchical and horizontal knowledge structures in his recognition of internal and external languages of description (which he labels  $L^1$  and  $L^2$ , respectively).  $L^1$  “refers to the syntax whereby a conceptual language is created” or how constituent concepts of a theory are interrelated, and  $L^2$  “refers to the syntax whereby the internal language can describe something other than itself” (2000, p. 132) or how a theory’s concepts are related to referents. Knowledge structures with a strong internal grammar ( $L^1$ ) have concepts that are tightly interrelated; in hierarchical knowledge structures, this facilitates the deployment of a strong external grammar ( $L^2$ ) whereby concepts are related to data in relatively unambiguous ways. Muller (2007) elaborates on these ideas, focussing on how knowledge structures progress (Muller, 2000, 2011; Moore and Muller, 2002). He introduces the term “verticality” to focus on how internal grammar develops—via ever more general propositions accounting for a broader range of data (more verticality) or the addition of new incommensurable languages of description (less verticality). He introduces the term “grammaticality” to focus on how knowledge structures manage data—via testable hypotheses about a restricted set of referents (strong grammaticality) or via readings of a less restricted set of referents that are hard to disconfirm (weak grammaticality). An outline sketch of these ideas is presented in Figure 2, including a rough positioning of canonical knowledge structures along a hierarchical/horizontal knowledge structure cline. Seen in these terms, Bateman’s ambitions for Multimodal Studies involve strengthening internal and external grammars of description so that the field can progress via what Bateman (2020a, p. 71) refers to as “explanatory sophistication” based on “worldly corroboration.”<sup>1</sup>

As far as grammaticality is concerned, Bateman (2021, p. 302–303) draws attention to Maton and Chen’s (2016) discussion and exemplification of mediating languages of description and external ones (termed  $L^{1.5}$  and  $L^2$ , respectively). Mediating languages are

designed to be more general and less data-specific than external languages. In SFL, for example, mediating languages comprise what are generally referred to as “descriptive motifs and generalizations” (Matthiessen, 2004)—i.e., general categories such as transitivity, modality, or tense (often presented as complementarities such as transitivity/ergativity, modality/assessment, or tense/aspect). These help a linguist approach the description of the grammar of a language with relatively “soft eyes” before locking into a more specific description of the data to hand. What ends up counting as  $L^1$ ,  $L^{1.5}$ , and  $L^2$  is itself a process (Martin et al., 2020a, 2023), unfolding over time, as  $L^{1.5}$  motifs and generalizations are promoted to  $L^1$  status or  $L^1$  concepts are demoted to mediating  $L^{1.5}$  language status (or perhaps relegated to  $L^2$  external grammar). We focus more specifically on this process when we consider the evolution of SFS from SFL below.

### 3 SFL and SFS (internal and external grammars)

SFL itself comprises a number of different languages of description, as reflected in the Routledge and Cambridge handbooks (Bartlett and O’Grady, 2017; Thompson et al., 2019). Here, we will assume the model developed by Martin (1992, 2010, 2014), which is the one that has most strongly influenced Bateman (e.g., Bateman, 1998, 2008, 2020b)—hereafter referred to simply as SFL. In relation to other social sciences, SFL has a strong internal grammar. Following Saussure (1916/1959), it treats language as a system of signs. Following Firth (1957), it takes the complementarity of paradigmatic and syntagmatic relations as fundamental. Following Halliday (1966, 1992), it skews this complementarity, privileging system over structure. This axial orientation underpins all language description, resulting in external grammar which formalizes value in networks of options realized in structure (Martin et al., 2013). Over time, SFL’s internal grammar has expanded to include the notion of hierarchy—i.e., realization (levels of abstraction), instantiation (a cline of sub-potentialisation/generalization), and individuation (a scale of allocation/affiliation). Of these, realization has the strongest grammar as systems in system networks bundle together in relation to the size of the structural unit realizing options (rank), the ideational, interpersonal, or textual meaning and corresponding types of particulate, prosodic, or periodic structure involved (metafunction) and the level of abstraction (phonology/graphology/signology, lexicogrammar, and discourse semantics). A synoptic overview of these dimensions (following Martin, 2010) is presented in Figure 3 (using an English MOOD system to represent axis). Of these, both instantiation and individuation are underarticulated compared with realization and constitute major challenges for future research.

To this compilation, I will add five elaborations that bear on the discussion. First, **form vs. substance**. As clarified by Martin et al. (2013), the register of SFL at stake here follows Saussure (1916/1959) and Hjelmslev (1961) in treating language as form, not substance. This means that phonetics is not treated as a stratum of language in its own right. Rather it is a region in Bernstein’s terms (interfacing with practices such as speech

<sup>1</sup> It is important to acknowledge that this modeling presents a “deficit” view of the humanities, a point which needs to be redressed but is unfortunately beyond the scope of this paper.



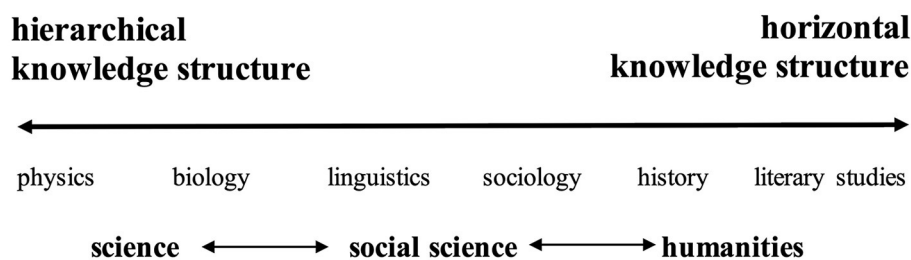


FIGURE 2  
Knowledge structures (vertical discourse).

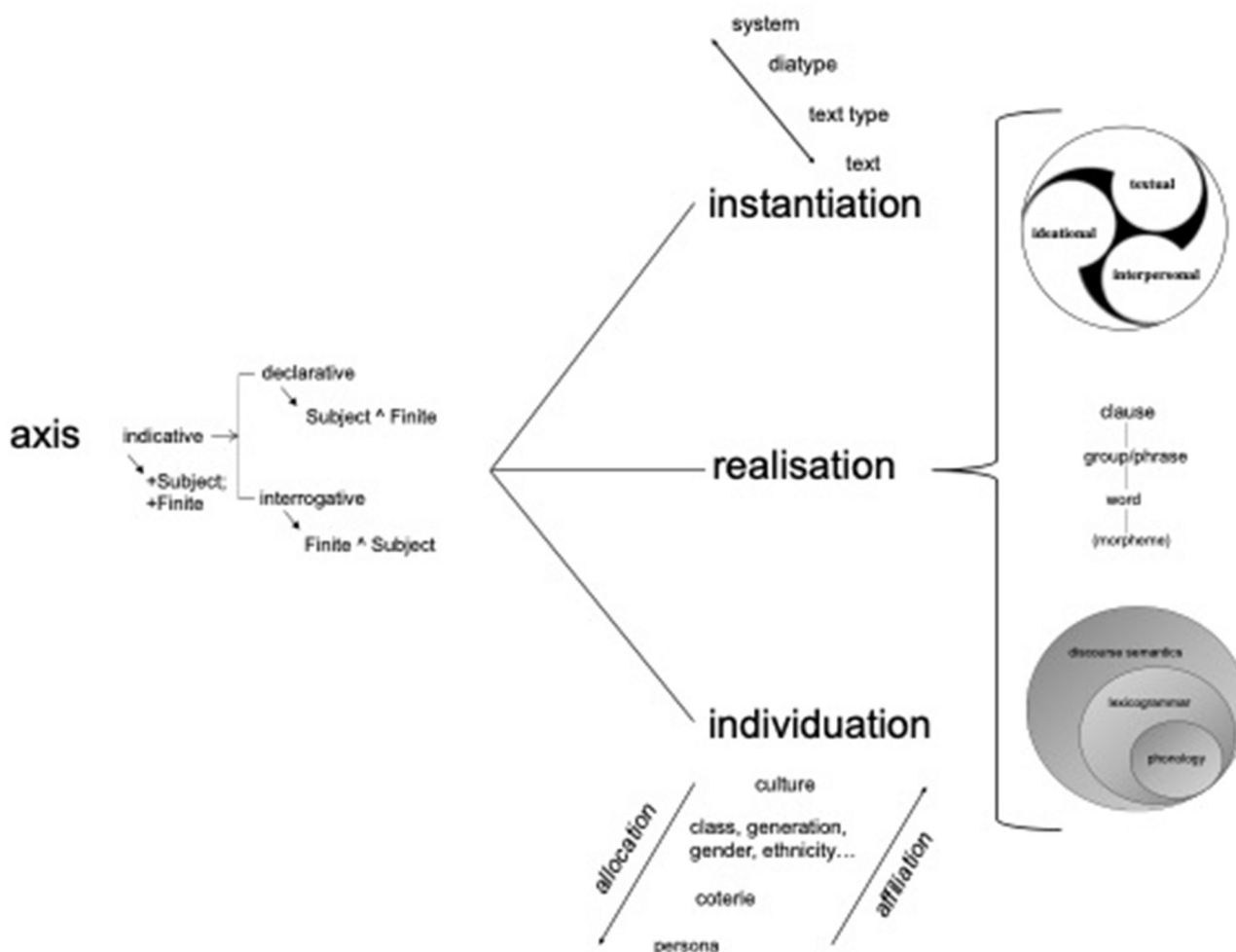


FIGURE 3  
SFL internal grammar ( $L^1$ ). From an axial perspective, for some languages, there is no need to distinguish word and morpheme ranks (since there is no word structure—i.e., no words consisting of more than one morpheme); parentheses make room for this variability at the bottom of the rank scale in this figure.

recognition or speech pathology), drawing on physics (acoustic phonetics) and neuro/biology (articulatory phonetics)—and thus deploying a set of internal and external grammars very different to those employed by linguistics proper (stronger grammars in fact). This is not to deny the relevance of phonetics and phonology to one another (linguistics students are generally trained in both) but simply to acknowledge the very different

knowledge structures involved in the description of form as opposed to substance.<sup>2</sup>

<sup>2</sup> Note that in more recent work, Halliday and Matthiessen (2014, p. 26) adopt a position similar to Bateman's, treating phonetics as a stratum of language. I will not pursue an argument with them here.

Second, **structure and syntagm**. As far as the description of grammatical “form” is concerned, SFL does not restrict its description to what [Whorf \(1945\)](#) called phenotypes—i.e., single or multi-segment syntagms consisting (for grammar) of classes of morpheme, word, group/phrase, or clause. To develop rich meaning-making grammatical descriptions, SFL is also inspired by Whorf’s notion of cryptotypes. [Halliday and Matthiessen \(2014\)](#) distinction between Epithet and Classifier in English can be used to illustrate this point here. From the perspective of system, English nominal groups make a distinction between describing and classifying. Both types of nominal groups can be realized by the same syntagm (i.e., determiner ^ adjective ^ noun), but a covert distinction can be uncovered by asking whether the adjective in the relevant syntagm is gradable or not. Describing adjectives are gradable (*a really lovely film*), whereas classifying adjectives are not (*\*a very Korean film*).<sup>3</sup> Accordingly, the same syntagm is assigned different structures, as in examples (1) and (2) below. SFL grammar descriptions in other words are not simply a catalog of syntagms; they build function structures on top of syntagms to reflect the meaning-making valeur at stake.

(1)

	<i>a</i>	<i>Korean</i>	<i>Film</i>
Structure	Deictic	Classifier	Thing
Syntagm	determiner	adjective	noun

(2)

	<i>a</i>	<i>Lovely</i>	<i>Film</i>
Structure	Deictic	Epithet	Thing
Syntagm	determiner	adjective	noun

This means, for example, in relation to an SFS description of some horizontally polarized images that an optional information value system can be set up realized by the function structure Given ^ New, without making the claim that all horizontally polarized imagic syntagms in fact realize this system. Relevant options are extended from [Kress and van Leeuwen \(2021, p. 216–217\)](#) in [Figure 4](#). The name of the system is INFORMATION VALUE; it is an optional system; if the feature [newsy] is selected, then the structural functions Given and New are present, in the sequence Given followed by New (with Given realized to the left and New to the right). This formalization makes no claims about all horizontally polarized systems; it simply positions [newsy] ones as having a Given ^ New structure realized by a horizontally polarized imagic syntagm. Note in passing that this is perhaps too generous a reading of Kress and van Leeuwen’s often criticized account of information structure in polarized images, but my point here is that SFS need not fall foul of their apparent overgeneralisations.

This approach to axis (i.e., paradigmatic and syntagmatic relations) lies at the heart of SFL/SFS descriptions of semiosis; [Martin et al. \(2013\)](#) provide a basic introduction. As in [Figures 5, 6](#), it privileges the formalization of paradigmatic relations as the basic organizing principle of descriptions and derives structural realizations from choices for meaning. The approach leaves open

the question of whether systems need to be set up to generalize the syntagms available for realizing function structures in a given semiotic system—with reference to [Figure 4](#), for example, opposing all polarized images to non-polarized ones, and if polarized, all horizontally opposed images to all vertically opposed ones. This could be important if polarized images are used to realize different function structures (and thus different meanings) for a given semiotic system. Arrows (as opposed to lines) and grids (as opposed to vertical or horizontal alignment) are good examples of imagic syntagms that arguably need generalization in such terms—since arrows are used to realize motion or links, for example (not to mention the system network specific uses of arrows in [Figure 4](#)), and grids can realize cross-classification (as in linguists’ paradigms) or momented activity (as in comics), for example (not to mention culturally specific arrangements such as that organizing Shirley Purdie’s remarkable artwork Goowoolem Gijam “Gija plants” which features at the Museum of Contemporary Art in Sydney).<sup>4</sup>

[Martin and Unsworth \(2024\)](#) take this step in their work on secondary school science infographics, drawing directly on work by [Hiippala et al. \(2021\)](#), [Hiippala \(2023\)](#). Martin and Unsworth’s network for MACRO-GROUPING is presented in [Figure 5](#). Therein, a square bracket means “or” (as shown in [Figures 2, 3](#)), a slanted square bracket indicates a cline, a brace means “and,” and a combination of brace and square bracket means “and/or.” So for the CO-TEXT systems, we have the option of including a text block or not, and if we choose to do so, we can include a caption or an interpolation or both. To follow one path in the DESIGN system, if we choose line, it can be more or less vertical or horizontal or both (in the latter case we end up with a grid).

If this syntagm oriented step is taken, then an analysis dedicated to such regularities of form can be established (e.g., [Caple, 2013, 2022](#) on BALANCE systems for images), and some kind of stratification of “meaning” and “form” can potentially be brought into the description (as in [He’s, 2021](#) work on animations). There is nothing in the knowledge structure of SFS, as informed by SFL, blocking stratified generalizations of this kind.

A related point about knowledge structure and SFS can be made in relation to “etics” and materiality. As [van Leeuwen \(1999, 2011\)](#) shows through his work on parametric systems for sound and color, axis can be used to formalize descriptions that cover the material oppositions which afford traces of function structures and syntagms of the kind introduced above. I would hesitate to refer to these systems as a stratum of language or any other semiotic system since clearly something other than semiotic internal and external grammars inform their description (the binary scaled simultaneous nature of “parametric” systems reflects exactly this point); in Hjelmslev’s terms, we are dealing with substance, not form. As emphasized above in relation to phonetics and phonology, this is not to suggest that work on materiality is not relevant to semiosis. It is simply to restrict stratification to cases where we have bundles of interdependent systems at different levels of abstraction

3 Unless we are in fact using *Korean* as an Epithet, describing characteristics of the genre.

4 This grid features 72 45 cm by 45 cm paintings of Kimberley flora, in four rows of 18 panels each, arranged top down as rows of taller plants and trees, then smaller plants and shrubs, and then plants from in or around water and ground dwelling plants—an Indigenous arrangement which could only be abducted by viewers very familiar with Gija culture.

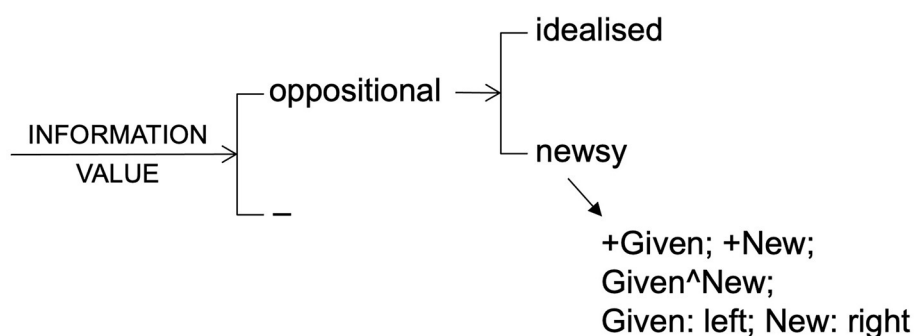


FIGURE 4

INFORMATION VALUE (system and structure). The downward slanted arrow in this diagram specifies the structural consequences of the feature [newsy]—namely, insert the function Given, insert the function New, sequence Given before New, realize Given through a left imagic block, and realize New through a right imagic block. The fourth edition of *Reading Images* (Kress and van Leeuwen, 2021, p. 217) in fact uses images to specify the realization of imagic functions.

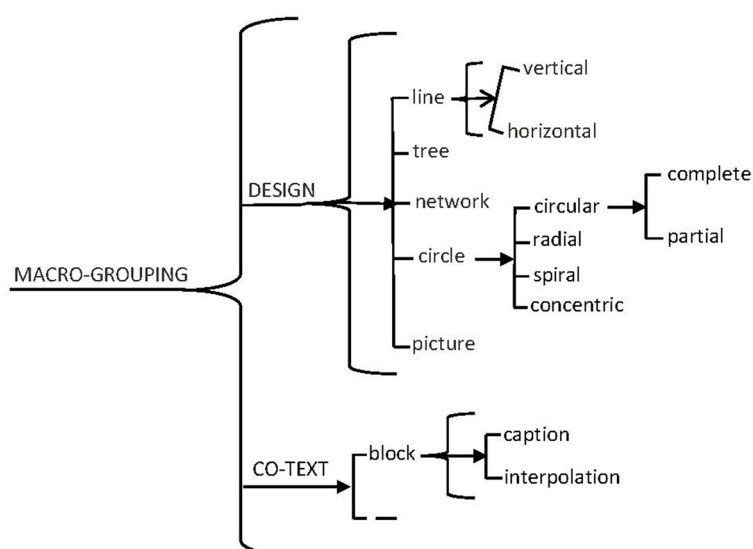


FIGURE 5

MACRO-GROUPING systems (Martin and Unsworth, 2024, p. 107).

(in a pattern of patterns relationship referred to by Lemke, 1984 as metaredundancy). The relation of “emics” to “etics” is not strictly speaking a pattern of this kind.

Third, **axis**. In SFL, other dimensions of internal grammar are all articulated in relation to axis (Martin et al., 2013). The concepts of rank, metafunction, and strata, in other words, are all based on bundles of interdependent features (organized with respect to constituency, type of meaning, or level of abstraction, respectively). Instantiation has to do with the manifestation of system in text and the generalization of instances as system, over time—as texts unfold (logogenesis), as speakers mature (ontogenesis), or as languages evolve (phylogenesis). Individuation has to do with the allocation of systems to members of a culture and their use of those systems to affiliate in social groups—once again, over time. Take away axis (e.g., O’Toole, 1994) and you may arguably be left with one kind of functional theory or another but not **Systemic** Functional

Linguistics or **Systemic** Functional Semiotics as the disciplines are being construed here.

As suggested by Martin (2011a), in multimodal studies which take axis as fundamental (i.e., SFS), it is critical not to make a priori assumptions about how systems will enter into interdependency relations with one another. Depending on the semiotic system in question, constituency (rank), kind of meaning and type of structure (metafunction), and level of abstraction (stratification) may shape external grammar ( $L^2$ ) but may not. Looking across the SFS studies surveyed in Table 1, the constant  $L^1$  notion is axis. Accordingly in SFS rank, metafunction and strata are clearly better positioned as mediating  $L^{1.5}$  notions—possibly shaping the description ( $L^2$ ), possibly not.

Fourth, **delicacy**. Recognition of mediating languages of description ( $L^{1.5}$ s) carries with it the idea that the relation between  $L^1$  and  $L^2$  can be treated as a cline. SFL’s approach to axis is well

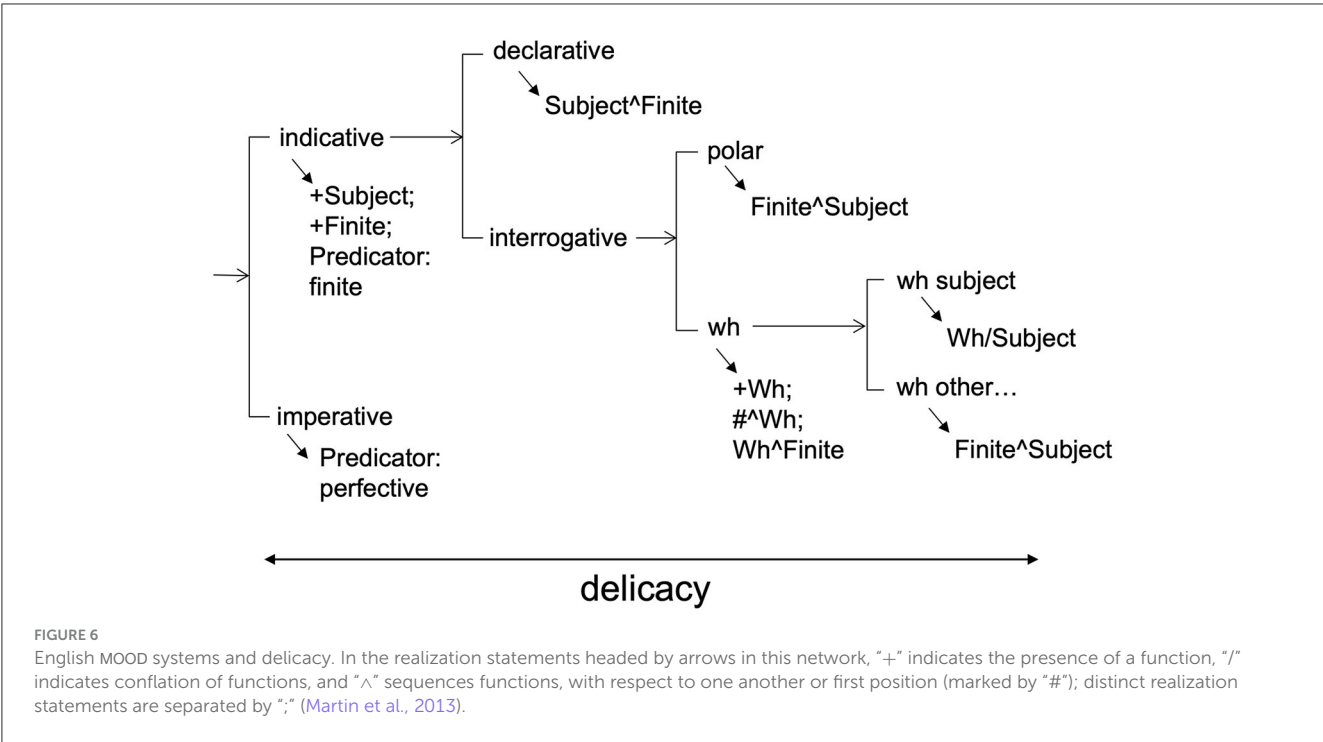


TABLE 1 SFS studies in relation to axis, rank, metafunction, and strata.

	Semiosis in focus	Axis	Rank	Metafunction	Strata
<i>Reading images</i> (Kress and van Leeuwen, 2021)	Images	+		representation, interaction, validity, composition	
<i>Photojournalism</i> (Caple, 2013)	News images	+		balance	
<i>Reading Visual Narratives</i> (Painter et al., 2013)	Picture book images	+		ideational, interpersonal, textual	
<i>Discourse of Physics</i> (Doran, 2018)	Mathematical symbolism	+	+	textual, logical, operational	+
“Toward a stratified metafunctional model...” (He, 2021)	Science animations	+		ideational, interpersonal, textual	+
<i>Modeling Paralanguage...</i> (Ngo et al., 2022)	Body language	+		ideational, interpersonal, textual	
<i>Reading Images for Knowledge Building</i> (Martin and Unsworth, 2024)	Science infographics	+		ideational, interpersonal, textual	
<i>Emoji and Social Media</i> (Zappavigna and Logi, 2024)	Emoji	+		ideational, interpersonal, textual	
<i>Multimodal Knowledge Building...Chemistry...</i> (Yu, forthcoming)	Chemical symbolism	+	+	experiential, logical	

adapted to a conception of this kind since systems are arranged from more general to more specific along a cline referred to as delicacy. Thus, in Figure 6, for example, indicative clauses are more general than interrogative ones, which are in turn more general than wh ones and so on. This makes it possible to be more and less specific about what gets treated as  $L^1$ ,  $L^{1.5}$ , and  $L^2$ —perhaps treating the [indicative/imperative] opposition as  $L^{1.5}$ , but remaining agnostic about more delicate options, pending construction of  $L^2$  (i.e., a specific language’s grammar of MOOD). It is important to keep in mind in relation to this point that positioning more or less general systems as  $L^{1.5}$  can be done without

making any claims at all about how such systems are realized in structure. Commitment to structural realizations of mood options needs to be withheld for  $L^2$ , since structures realizing mood vary considerably across languages (Martin, 2018; Martin et al., 2021).

Fifth, **context**. In SFL, context is treated as form (Figure 7), not substance (as connotative semiotics in Hjelmslev’s terms). In the model of SFL assumed here, context is stratified as register and genre (Martin, 1992; Martin and Rose, 2008), and register is modeled metafunctionally in proportion to its realization in language—i.e., field is to ideational meaning, as tenor is to interpersonal meaning, and as mode is to textual meaning. This

move allows context strata to be treated as resources for making meaning and modeled axially, comparably to language ones (Doran and Martin, 2021; Doran et al., 2024; Martin and Doran, in press). These context dimensions of SFL's  $L^1$  are less well articulated and more controversial than language ones—probably because of their level of abstraction and concomitant realization as patterns of language patterns in addition to a lack of clarity in argumentation as far as the distinction between realization and instantiation is concerned.

We are now in a position to compare SFS with Bateman's proposals for the basic  $L^1$  of what he refers to as a semiotic mode (a semiotic system in SFS).<sup>5</sup> In doing so, we need to keep in mind the potential of SFS's  $L^1$  in relation to extant descriptions of semiotic modes ( $L^2$ ). External grammar ( $L^2$ ) can be critiqued and revised in ways that do not call for renovations or reconstructions of  $L^1$  or  $L^{1.5}$ . Some shortcomings, in other words, are more serious than others—as we shall flag below.

## 4 Bateman's model of multimodality

SFL's stratified model of language and context, as presented in Figure 7, provides a useful point of departure for comparison with Bateman's diagrammatic illustrations of his modeling of semiotic modes. Bateman et al. (2017, p. 117) introduce the diagram in Figure 8, consisting of three strata—"the material substrate or dimension, the technical features organized along several axes of descriptions (abbreviated as 'form'), and the level of discourse semantics." Compared to SFS, this treats the "etics" of materiality as a stratum proper and groups it together with the stratum of form—the two together realizing discourse semantics.

Bateman's approach to discourse semantics is a dynamic one, whereby what he calls form is imbued with meaning as texts unfold. As outlined in Bateman (2022b, p. 69), forms are not treated as already themselves contextually meaningful. He continues:

"...meanings are only mediated by the application of the *discourse semantics* of the semiotic mode... Thus, as an example, whereas the often used classification of graphical resources set out in Kress and van Leeuwen (2006 [1996]: 59–68) might classify graphical 'arrows' as 'narrative processes' (by virtue of their directionality as vectors), from the perspective of the approach adopted here this conflates two semiotic strata of description: the formal level at which visual properties of connection and directedness properly reside, and the discourse semantic level at which, under certain circumstances, it may be possible to abduce that the graphical connective is serving a 'narrative purpose' (but then, in other circumstances, it may not be). ... meaning-making using semiotic modes is best characterised as discourse 'unfolding' and it is this that offers a higher 'unity' to any material regularities exhibited."

This characterization of meaning making in semiotic modes is helpfully reviewed in Bateman (2020b) in relation to Martin (1992) model of discourse semantics. Seen in SFS terms, Bateman's

discourse semantics is strongly focussed on instantiation—logogenesis in particular. From this perspective, meaning is only ever something that can be abduced in relation to co-text and context as texts build meaning—rolling out a snowball of semiosis as they unfold. So what Bateman treats as a stratum called discourse semantics, SFS would interpret from the hierarchy of instantiation, not realization. SFL's discourse semantic stratum, along with lexicogrammar and phonology, would all be treated as form in Bateman's modeling (i.e., "technical features organized along several axes of descriptions"; Bateman et al., 2017, p. 117). Of these, Martin's notion of covariate structure (e.g., Martin, 2015) comes closest to Bateman's conception of discourse semantics—since covariate structures instantiate discourse semantic systems through a process of abducing relations of indefinite extent as texts unfold. Martin's stratum of discourse semantics would have to be interpreted in Bateman's modeling as proposals for the systems of relations that can be so abduced.<sup>6</sup>

In at least one articulation of Bateman's model (Bateman and Schmidt, 2012, p. 81), discourse semantics as well as form is presented as involving paradigmatic systems of choice and syntagmatic organization imposing structure. This modeling informs Bateman (2007) work on semantic relations between shots in film (his "grande paradigmatique") and Tseng and Bateman (2011) description of filmic identification—with systems formalizing relations to be abduced. In later work, perhaps because of reservations about the synoptic, non-dynamic nature of system networks,<sup>7</sup> axis seems to be reserved for the stratum of form. In its place, at the level of discourse semantics, lists of relations, elaborated from Mann and Thompson (1986) Rhetorical Structure Theory (RST), regularly function as reservoirs of meaning to be abduced (e.g., Bateman, 2008; Hiippala, 2015).

As acknowledged in Section 3, compared to realization, instantiation is to date a still developing hierarchy in SFL. Suggestive explorations include Martin (2006) on intralingual re-instantiation, de Souza (2013) on interlingual re-instantiation, Martin (2010) on coupling and commitment, Painter et al. (2013) on intermodal convergence (concurrence, resonance, and synchronicity), Martin and Matruglio (2013) on presence, Martin (2017) on mass, Martin and Doran (in press) on context as realization vs. instantiation, and Martin and Unsworth (2024) on syndromes of instantiation referred to as mass, presence (and association in (Martin and Doran, in press)). None of these approaches comes anywhere near the level of explicitness and detail underpinning work by Bateman and his colleagues (e.g., Bateman and Rhonhuis, 1997; Wildfeuer, 2021), inspired as it is by Asher and Lascarides' work on the logic of abduction (e.g. Lascarides and Asher, 1991, 1993; Asher and

6 Tellingly, in Martin (1992), realization statements are not provided for covariate structures realizing discourse semantic options—precisely because such structures have to be abduced as discourse semantic systems are instantiated in texts.

7 This problem is elaborated in relation to exchange structure and genre analysis in Martin (1985). Bateman (2020b) comments on the lack of progress in SFL as far as transcending axis and developing meaning building approaches to unfolding discourse are concerned.

5 As noted above, the term mode is used for a register variable in SFL, so the term semiotic system is preferred to semiotic mode in SFS.



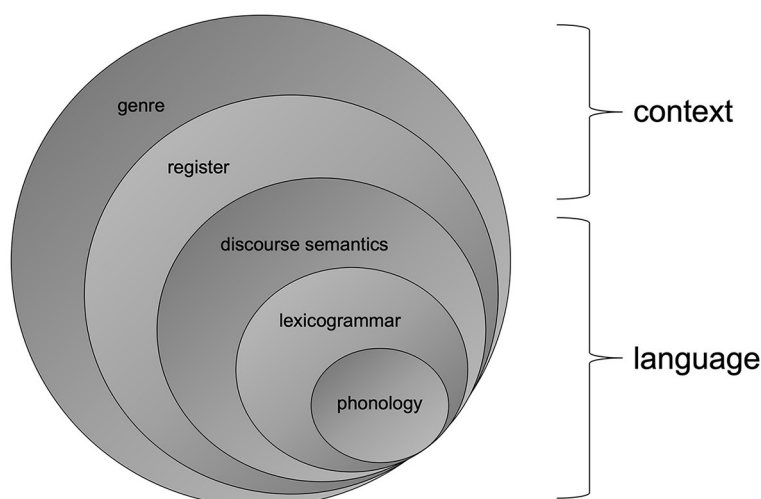


FIGURE 7  
Stratified model of language and context.

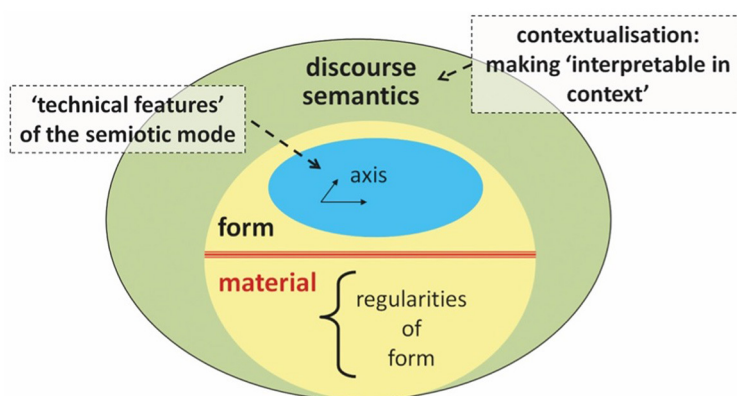


FIGURE 8  
Bateman et al. (2017) "abstract definition of a semiotic mode."

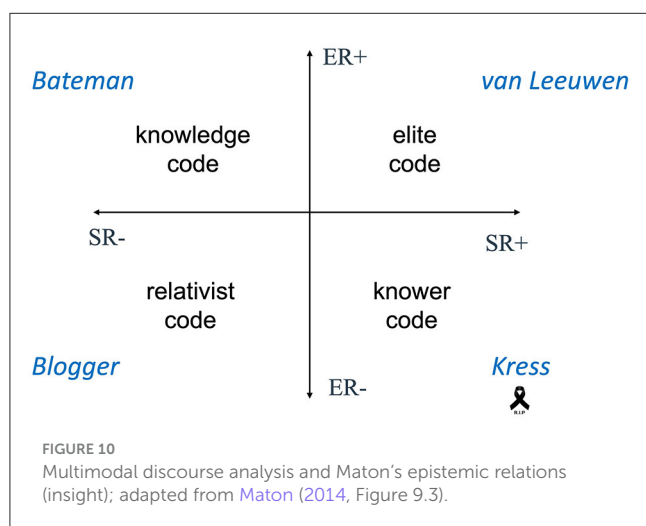
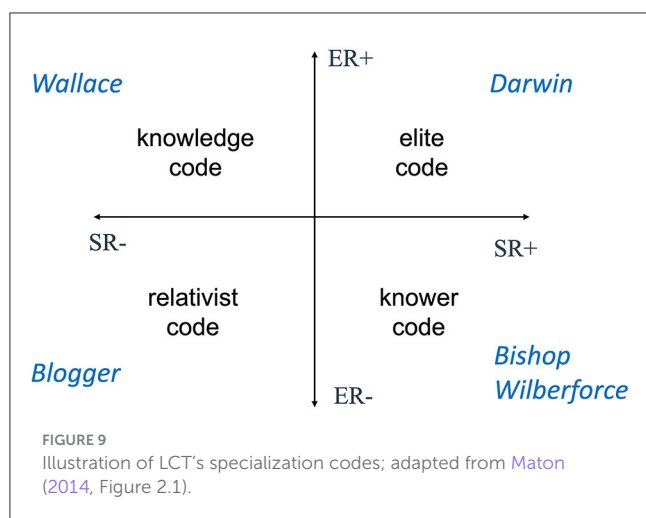
Lascarides, 2003). This conceptual shortcoming is clearly one area where SFL and SFS are certain to be positively influenced by Bateman's modeling of multimodality (discourse semantics in particular).

In addition to the three strata outlined in Figure 8, alternative L<sup>1</sup> imaging of Bateman's model includes a more abstract stratum comprising "context, social norms, and values" (Bateman and Schmidt, 2012, p. 81) or "context/register/situation" (Bateman, 2022b, p. 69), and in related work (e.g., Bateman, 2008, 2016), the notion of genre is brought into the picture. This work resonates with SFL work on modeling context as register and genre, and there is clearly room for ongoing collaboration in this area. That said, one weakness of some SFS work on semiotic systems such as that noted in Table 1 is the lack comprehensive treatments of register and genre—especially for descriptions clearly inspired by Kress and van Leeuwen (1990 and subsequent editions) grammar of images. Further work, modeled on the attention paid to field and genre in Doran's (2018) work on physics and mathematics or to field in Yu (forthcoming) work on chemistry, is clearly in order.

## 5 Knowledge and knowers

As foregrounded in Maton's Legitimation Code Theory (LCT), disciplinarity needs to be considered from the complementary perspectives of knowledge and knowers (Maton, 2014). Accordingly, LCT's legitimation code referred to as specialization takes into account both epistemic relations (between knowledge and what it describes) and social relations (between knowledge and who is producing it). Based on the relative strength of epistemic and social relations, LCT establishes a topology<sup>8</sup> of legitimation codes, with four principal "modalities"—knowledge codes (ER+, SR-) for which legitimacy depends on what you know; knower codes (ER-, SR+) where what matters is who you are; elite codes (ER+, SR+) where it matters both what you know and who you are; and relativist codes (ER-, SR-)

<sup>8</sup> The horizontal and vertical axes in the LCT derived topologies in Figures 9, 10 represent clines, as imaged by the double-headed arrows.



through which everyone's voice and knowledge is equally valid (Maton, 2014, p. 30–33). This framework is exemplified in Figure 9 with reference to speaking rights in 19th century debates about evolution. The key players were Watson (knowledge code), a middle class biologist who made a living by selling specimens to private collectors and museums, Darwin (elite code), a wealthy biologist who married into the Wedgewood family pottery fortune, and Bishop Wilberforce (knower code), a gifted public speaker and high ranking clergyman. They are positioned according to specialization in Figure 9 below. Tellingly, it is Darwin (not Watson) who comes readily to mind when we think of evolution, along perhaps with Bishop Wilberforce who played the role of Huxley's protagonist in the famous debate at a meeting of the British Science Association in 1860. To complete the picture, I have imagined a know-it-all nameless blogger (relativist code), opining about evolution in the 21st century.

Turning to multimodality, we might position Bateman as legitimized by a knowledge code, van Leeuwen (a professional jazz musician and filmmaker) by an elite code, someone

such as Lima<sup>9</sup> by a knower code, and our blogger (still blogging) by a relativist code.<sup>10</sup> Kress's positioning would have to be a more transitional one, beginning ER+/SR- with the publication of *Reading Images* (1990), but sliding toward ER-/SR+ thereafter—viz., publications such as *Literacy in the New Media Age* (2003) and *Multimodality: a social semiotic approach to contemporary communication* (2010) which typify his later work, rarely draw on any analysis at all from *Reading Images* and popularize his thinking. With his passing in 2019, the field of multimodality lost its most influential knower. This is significant because turning a region into a singular with a strong internal and external grammar requires a range of voices (including popularizing knowers), even if legitimation via a knowledge code is what protagonists such as Bateman ultimately have in mind.

Digging deeper into epistemic relations, Maton (2014, p. 175–177) distinguishes relations between knowledge practices and the part of the world they are oriented to (ontic relations, OR) and relations between knowledge practices (discursive relations, DR). Stronger ontic and discursive relations (OR+/DR+) establish a purist code, which emphasizes both the object of study and how it is studied. Weaker ontic relations but strong discursive relations (OR-/DR+) establish a doctrinal code, which legitimate a multiplicity of objects of study but foregrounds a particular way of studying them. Stronger ontic but weaker discursive relations (OR+/DR-) allow for a situational code, whereby a specific object of study is in focus but it can be approached from multiple points of view (or as Maton allows, no clear knowledge code at all). Weaker ontic and discursive relations (OR-/DR-) legitimate an approach, which is unlimited with respect to both what is studied and how. As suggested in Figure 10, a framework of this kind would position SFL as purist (since it studies language from one theoretical perspective), SFS as doctrinal (since it studies any semiotic system but always from the perspective of SFL informed theory), social semiotics (a la Kress and van Leeuwen, 2005) as DR-/OR- (since it encourages the study of multimodality from different points of view), and something such as information visualization as DR-/OR+ (since it focusses on graphic representations of complex data by whatever means afford a clear “synoptic” overview). Seen in these terms, one of Bateman's objectives is to re-orient the current trajectory of Multimodal Studies, which at some conferences seems to sprawl toward ever weaker discursive and ontic relations; he wants to shift its trajectory toward stronger discursive relations whatever its object of study (without, we might reiterate here, falling foul of the linguistic imperialism and predatory interdisciplinarity that SFS's doctrinal stance might be accused of).

The main message to take from Maton's work on specialization and epistemic relations is that disciplines involve both knowledge and knowers. Prescribing strong internal and external grammars

9 Lima is the only “multimodalist,” as far as I am aware, who has been called on to give a TED talk. His best-known publications (Lima, 2011, 2014, 2017) are popularisations.

10 I should perhaps emphasize here that codes do not ascribe value to speakers in different ways; they simply characterize the factors that legitimize a given voice.



for Multimodality Studies is not enough; a given field needs knowers as well. In this regard, SFS has the advantage of being able to recruit both knowledge and knowers from SFL, since in practice SFL informed discourse analysis cannot avoid bumping into multimodal texts and the technicality of SFL's internal and external grammar is already in play. Given Bateman's vision for Multimodal Studies, recruiting knowers from social semiotics is perhaps more of a challenge since its relatively weak internal and external grammar and its multidisciplinary stance make stronger grammar a harder sell. Work that draws on [Lascarides and Asher \(1991, 1993\)](#) to formalize the complexity of discourse semantic abduction seems certain to frighten large numbers of OR-/DR-multimodalists well away.

## 6 Fair play

In this paper, I have drawn on Bernstein and Maton's sociology of knowledge to explore SFL and SFS in relation to Bateman's vision for empirical multimodality research. There is of course much more to survey. In closing, let me just highlight three main points here.

First, there is the question of which theoretical dimension is privileged as fundamental. For SFL/SFS, this is axis; all other dimensions of the theory, including stratification, depend on a specific conception of paradigmatic relations underpinning and underpinned by syntagmatic ones. For Bateman, the fundamental dimension is stratification, further specified as materiality, form, and discourse semantics. So where SFL/SFS derives strata from axis depending on the interdependency of systems according to levels of abstraction in a particular semiotic system, Bateman's vision assumes three strata and in recent work uses axis to characterize just one of these (i.e., form). Related to this point is Bateman's treatment of materiality as a stratum, whereas in the model of SFL/SFS assumed here, it would be treated as "etic" substance and explored through knowledge structures that have evolved for the study of physical and biological reality (as opposed to those which have evolved for exploring semiotic reality, i.e., systems of meaning).

Second, and perhaps most crucially from Bateman's perspective, discourse semantics is approached from a dynamic perspective by Bateman—with form imbued with meaning through a process of abduction as texts unfold. This is compatible with SFL/SFS's approach to instantiation (logogenesis in particular) and its conception of covariate structure, but as noted above, SFL/SFS description ( $L^2$ ) has not caught up with theory ( $L^1$ ) as far as instantiation is concerned. Bateman's misgivings about SFL/SFS's many promissory notes in this regard are right on target.

Third, in SFL/SFS, key concepts are deployed across strata. One dimension, axis, is fractal; all strata, ranks, and metafunctions are explored axially, and systems of choice shape SFL/SFS's conception of hierarchy (realization, instantiation, and individuation). This axial orientation grounds decisions for a specific semiotic system—with respect to how many strata, how many ranks, and which metafunctions (if any) are presumed as  $L^1$ , suggested as  $L^{1.5}$  or described as  $L^2$  ([Martin et al., 2013](#)). Bateman's model is more modular in design, with distinct internal and external grammars proposed for each stratum (and for context, it would appear, once we move beyond his semiotic modes). The accessibility of work on materiality by Bateman et al. (i.e., their slices of canvas) contrasts

markedly with the technicality of their adoption of Lascarides and Asher's formalization of the logic of abduction. To be frank, it is clear to me that such an approach potentially formalizes the complexity of what is going on ideationally in logogenesis as far as the snowballing of meaning is concerned, but I am much less clear about how it manages this complexity for descriptive or applied purposes (especially once we scale up and move beyond the fragments of exemplificatory discourse used as illustrations of the approach). This may simply be a matter of unfamiliar technicality and the challenge it imposes on outsiders (such as myself), but it may be more than that. As I often tell my research students when they are feeling overwhelmed by the phenomena they are describing, there is a difference between documenting complexity and managing it. The job of internal and external grammars in any discipline is to manage complexity, not simply catalog it. I will leave the much needed discussion of this instantiation modeling crisis for another time (to another generation perhaps, who can come to our rescue in this regard).

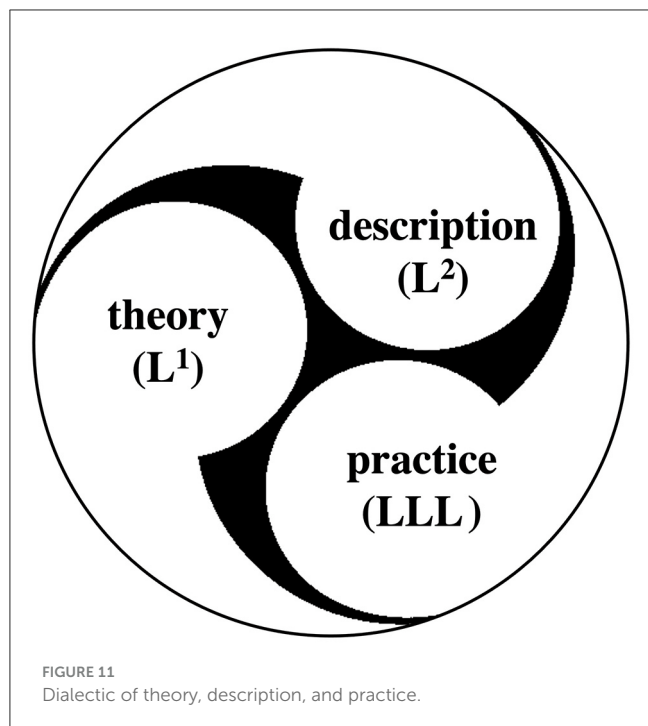
Overall, my comment would be that SFL/SFS's internal grammar is by and large compatible with Bateman's vision, even if its external grammar falls short insofar as extant descriptions of one semiotic system or another are concerned. This is hardly surprising since SFL has had a considerable influence on Bateman's thinking. Where differences arise, I think that by and large the models can learn from each other—provided suitable contexts for working together are formed.

Drawing on one knowledge structure to position two others as I have done is a challenging task, but it has the salutary advantage of drawing attention to the demands of coming to grips with the technicality of incommensurable  $L^1$ s—which can be forbidding given the every-worsening time constraints of academic life and the challenge of practical applications which demand solutions yesterday for what really needs to happen tomorrow. This takes me to my final point, applicability, which in my experience bears critically on what it takes to come to grips with an unfamiliar  $L^1$  and the  $L^2$  descriptions it affords.

[Halliday \(2008, p. 7\)](#) coins the term "applied linguistics" for the dialectic of theory, description, and practice informing his linguistic work. An outline of this problem-oriented perspective is outlined in [Figure 11](#) (with multiple  $L$ s allowing for the probability of a number of singulars influencing a given region of practice). This orientation to linguistics was an unusual one in the 20th century, the closest parallel being [Pike \(1988\)](#) tagmemics (viz., Pike's tagmemics and his "maxim" from 1988: "I wanted a theory that would allow one to live outside the office with the same philosophy one uses inside it").<sup>11</sup>

The challenge for SFS and Bateman's Multimodal Studies, as I see it, lies precisely in finding contexts of application (in educational semiotics, clinical semiotics, forensic semiotics, etc.), which foster a dialectic of theory, description, and practice. It is in these contexts, especially if knowledge workers share a politics in relation to a specific problem, that the challenge of incommensurable technicality can be overcome. This dialectic

11 From "Kenneth L. Pike Maxims"; <https://www.sil.org/about/klp/pike-maxims>. Pike's Christian motivations were of course very different from Halliday's Marxist ones ([Halliday, 2015](#)).



works best when theory and description provide complementary perspectives on the same data. I am not in other words talking about interdisciplinarity (i.e., “you do your bit, I’ll do mine”) but rather about transdisciplinarity (i.e., “this is how I see it, how about you?”). By approaching the same data from a different point of view, with a practical politically charged challenge in mind, theories can learn from one another—if they can make explicit what another theory is interested in and/or draw attention to regularities another theory/description has missed (Martin, 2011b; Maton et al., 2016; Martin et al., 2020b).<sup>12</sup>

It is for this reason that I have done what I can to encourage the evolution of SFS out of SFL—so it can “trespass” into conversations of this kind. Dialogue between SFS and Bateman’s vision for an empirical multimodality is likely to be far more productive than one involving SFL rather than SFS, precisely because SFS and Multimodal Studies can focus on the same data from complementary points of view. For me, then it is important to avoid dialogue in which linguists focus on language and multimodalists deal with everything else. Multimodal discourse needs to be the focus of all parties in the conversation.

<sup>12</sup> Note in this regard that I have strong reservations about the viability of what van Leeuwen (2005) refers to as an integrationist model of interdisciplinarity involving several knowledge structures which develop a ‘common metalanguage’. Singulars are incommensurable, and we have to always be on guard against borrowing terms or concepts from one knowledge structure into another and assuming they mean the same thing; they will not. What I am suggesting rather is that knowledge structures provoke one another and react by developing theory and description in their own terms. This is not of course to foreclose the possibility of a new Multimodal Studies singular emerging out of interdisciplinary or transdisciplinary work, which is on my reading what Bateman has in mind.

In saying this, I hope I am allaying fears about linguistic imperialism and have presented a less-than-predatory vision of SFS and SFL. One abiding concern I have is the “logophobia” generated by multimodality “knowers” through their by now rather dated proclamations of a new multimodality age and the striking absence of any language analysis to speak of in key multimodal conferences and publications. Language needs to be part of the picture, if multimodalists are serious about what is going on.

The title of my paper is of course a provocative one—which I propose to suggest that we need to be careful about how we weigh up the advantages and disadvantages of one large encompassing theory such as SFS (my “intradisciplinarity”) in relation to the advantages and disadvantages of the more generally celebrated interdisciplinary approach. This provocation is licensed in this context I hope by the incisive provocations marking Bateman’s interventions over many decades—calling out what needs calling out and challenging thinkers to think some more. Whenever my students embark on multimodal research, I warn them, “Read Bateman; he’s someone that is truly serious about what is going on.” For all this, John, my sincere thanks, many times over; and best wishes for the next phase of your brilliant career.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

## Author contributions

JM: Writing – original draft.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

- Asher, N., and Lascarides, A. (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.
- Bartlett, T., and O'Grady, G. (eds) (2017). *The Routledge Handbook of Systemic Functional Linguistics*. London: Routledge. doi: 10.4324/9781315413891
- Bateman, J. A. (1998). *James R. Martin's English Text: System and Structure. Functions of Language*. Amsterdam, 213–247. doi: 10.1075/fol.5.2.06bat
- Bateman, J. A. (2007). Towards a grande paradigmatique of film: Christian Metz reloaded. *Semiotica* 167, 13–64. doi: 10.1515/SEM.2007.070
- Bateman, J. A. (2008). *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*. London: Palgrave Macmillan. doi: 10.1057/9780230582323
- Bateman, J. A. (2016). "Methodological and theoretical issues in multimodality," in *Handbuch Sprache im multimodalen Kontext*, eds N.-M. Klug, and H. Stökl (Berlin: De Gruyter), 36–74. doi: 10.1515/9783110296099-003
- Bateman, J. A. (2020a). "Afterword: legitimating multimodality," in *Multimodality: Disciplinary Thoughts and the Challenge of Diversity*, eds J. Wildfeuer, J. Pflaeging, J. A. Bateman, O. Serizov, and C.-I. Tseng (Berlin: De Gruyter), 297–316. doi: 10.1515/9783110608694-013
- Bateman, J. A. (2020b). "The foundational role of discourse semantics beyond language," in *Discourses of Hope and Reconciliation: On J R Martin's Contribution to Systemic Functional Linguistics*, eds M. Zappavigna, and S. Dreyfus (London: Bloomsbury), 39–55. doi: 10.5040/9781350116092.ch-003
- Bateman, J. A. (2021). "Dimensions of materiality: towards an external language of description for empirical multimodality research," in *Empirical Multimodality Research: Methods, Evaluations, Implications*, eds J. Pflaeging, J. A. Bateman, and J. Wildfeuer (Berlin: De Gruyter), 36–63. doi: 10.1515/9783110725001-002
- Bateman, J. A. (2022a). Multimodality, where next? – some meta-methodological considerations. *Multimodality Soc.* 2, 41–63. doi: 10.1177/26349795211073043
- Bateman, J. A. (2022b). Growing theory for practice: empirical multimodality beyond the case study. *Multimodal Commun.* 11, 63–74. doi: 10.1515/mc-2021-0006
- Bateman, J. A., and Rhonhuis, K. J. (1997). Coherence relations: towards a general specification. *Discourse Processes* 24, 3–49. doi: 10.1080/01638539709545006
- Bateman, J. A., and Schmidt, K.-H. (2012). *Multimodal Film Analysis: How Films Mean*. London: Routledge. doi: 10.4324/9780203128220
- Bateman, J. A., Wildfeuer, J., and Hiippala, T. (2017). *Multimodality: Foundations, Research and Analysis – A Problem-Oriented Introduction*. Leiden: Brill. doi: 10.1515/9783110479898
- Bernstein, B. (1996). *The Structuring of Pedagogic Discourse*. (Class, codes and control Volume IV). London: Routledge.
- Bernstein, B. (1999). Vertical and horizontal discourse: an essay. *Br. J. Sociol. Educ.* 20, 157–173. doi: 10.1080/01425699995380
- Bernstein, B. (2000). *Pedagogy, Symbolic Control, and Identity: Theory, Research, Critique*, revised ed. Oxford: Rowman and Littlefield.
- Caple, H. (2013). *Photojournalism: A Social Semiotic Approach*. London: Palgrave MacMillan. doi: 10.1057/9781137314901
- Caple, H. (2022). "A balancing act: theorising compositional choices in photographs," in *Applicable Linguistics and Social Semiotics: Developing Theory from Practice*, eds D. Caldwell, J. S. Knox, and J. R. Martin (London: Bloomsbury), 41–54. doi: 10.5040/9781350109322.ch-2
- de Souza, L. M. F. (2013). Interlingual re-instantiation – a new systemic functional perspective on translation. *Text Talk* 33, 575–594. doi: 10.1515/text-2013-0026
- Doran, Y. J. (2018). *The Discourse of Physics: Building Knowledge through Language, Mathematics and Image*. London: Routledge. doi: 10.4324/9781315181134
- Doran, Y. J., and Martin, J. R. (2021). "Field relations: understanding scientific explanations," in *Teaching Science: Knowledge, Language, Pedagogy*, eds K. Maton, J. R. Martin, and Y. J. Doran (London: Routledge), 105–133. doi: 10.4324/9781351129282-7
- Doran, Y. J., Martin, J. R., and Zappavigna, M. (2024). *Negotiating Social Relations: (A Systemic Functional Perspective on) Tenor Resources in English*. London: Equinox. in preparation.
- Firth, J. R. (1957). *A Synopsis of Linguistic Theory, 1930-1955. Studies in Linguistic Analysis (Special volume of the Philological Society)*. London: Blackwell, 1–31.
- Halliday, M. A. K. (1966). Some notes on deep grammar. *J. Linguist.* 2, 57–67. doi: 10.1017/S0022226700001328
- Halliday, M. A. K. (1992). "A systemic interpretation of Peking syllable finals," in *Studies in Systemic Phonology*, ed. P. Tench (London: Pinter), 98–1121.
- Halliday, M. A. K. (2008). "Working with meaning: towards an applicable linguistics," in *Meaning in Context: Implementing Intelligent Applications of Language Studies*, ed. J. J. Webster (London: Continuum), 7–23.
- Halliday, M. A. K. (2015). "The influence of Marxism," in *The Bloomsbury Companion to M.A.K. Halliday* (New York, NY: Bloomsbury Academic), ed. J. J. Webster, 94–100.
- Halliday, M. A. K., and Matthiessen, C. M. I. M. (2014). *Halliday's Introduction to Functional Grammar*, 4th ed. London: Routledge. doi: 10.4324/9780203783771
- He, Y. (2021). Towards a stratified multimodal model of animation. *Semiotica* 239, 1–35. doi: 10.1515/sem-2019-0078
- Hiippala, T. (2015). *The Structure of Multimodal Documents: An Empirical Approach*. London: Routledge. doi: 10.4324/9781315740454
- Hiippala, T. (2023). Corpus-based insights into multimodality and genre in primary school science diagrams. *Vis. Commun.* doi: 10.1177/14703572231161829. [Epub ahead of print].
- Hiippala, T., Alikhani, M., Haverinen, J., Kalliokoski, T., Logacheva, E., Orekhova, S., et al. (2021). AI2D-RST: a multimodal corpus of 1000 primary school science diagrams. *Lang. Resour. Eval.* 55, 661–688 doi: 10.1007/s10579-020-09517-1
- Hjelmslev, L. (1961). *Prolegomena to a Theory of Language*. Madison, WI: University of Wisconsin Press.
- Kress, G. (2003). *Literacy in the New Media Age*. London: Routledge. doi: 10.4324/9780203299234
- Kress, G. (2010). *Multimodality: A Social Semiotic Approach to Contemporary Communication*. London: Routledge.
- Kress, G. (2015). Semiotic work: applied linguistics and a social semiotic account of multimodality. *AILA Rev.* 28, 49071. doi: 10.1075/aila.28.03kre
- Kress, G., and van Leeuwen, T. (1990). *Reading Images*. Geelong: Deakin University Press.
- Kress, G., and van Leeuwen, T. (2005). *Introducing Social Semiotics*. London: Routledge.
- Kress, G., and van Leeuwen, T. (2021). *Reading Images*. Geelong: Deakin University Press. doi: 10.4324/9781003099857
- Lascarides, A., and Asher, N. (1991). "Discourse relations and defeasible knowledge," in *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (Morristown, NJ: Association for Computational Linguistics)*, 55–63. doi: 10.3115/981344.981352
- Lascarides, A., and Asher, N. (1993). Temporal interpretation, discourse relations, and common sense entailment. *Linguist. Philos.* 16, 437–495. doi: 10.1007/BF00986208
- Lemke, J. (1984). *Semiotics and Education*. Toronto, ON: Toronto Semiotic Circle (Monographs, Working Papers and Publications 2).
- Lima, M. (2011). *Visual Complexity: Mapping Patterns of Information*. New York, NY: Princeton Architectural Press.
- Lima, M. (2014). *The Book of Trees: Visualizing Branches of Knowledge*. New York, NY: Princeton Architectural Press.
- Lima, M. (2017). *The Book of Circles: Visualizing Spheres of Knowledge*. New York, NY: Princeton Architectural Press.
- Mann, W. C., and Thompson, S. A. (1986). Relational propositions in discourse. *Discourse Processes* 9, 57–90. doi: 10.1080/01638538609544632
- Martin, J. R. (1985). "Process and text: two aspects of human semiosis," in *Systemic Perspectives on Discourse: selected theoretical papers from the 9th International Systemic Workshop*, eds J. D. Benson, and W. S. Greaves (Norwood, NJ: Ablex) (Advances in Discourse Processes 15), 248–274.
- Martin, J. R. (1992). *English Text: System and Structure*. Amsterdam: Benjamins. doi: 10.1075/z.59
- Martin, J. R. (2006). Genre, ideology and intertextuality: a systemic functional perspective. *Linguist. Hum. Sci.* 2, 275–298. doi: 10.1558/lhs.v2i2.275298
- Martin, J. R. (2010). "Semantic variation: modelling system, text and affiliation in social semiosis," in *New Discourse on Language: Functional Perspectives on Multimodality, Identity and Affiliation*, eds M. Bednarek, and J. R. Martin (London: Continuum), 1–34.
- Martin, J. R. (2011a). "Multimodal semiotics: theoretical challenges," in *Semiotic Margins: Meaning in Multimodalities*, S. Dreyfus, S. Hood, and M. Stenglin (London: Continuum), 243–270.
- Martin, J. R. (2011b). "Bridging troubled waters: interdisciplinarity and what makes it stick," in *Disciplinary: Functional Linguistic and Sociological Perspectives*, eds F. Christie and K. Maton (London: Continuum), 35–61.
- Martin, J. R. (2014). Evolving systemic functional linguistics: beyond the clause. *Funct. Linguist.* 1. doi: 10.1186/2196-419X-1-3
- Martin, J. R. (2015). Meaning beyond the clause: co-textual relations. *Linguist. Human Sci.* 11, 203–235. doi: 10.1558/lhs.34711

- Martin, J. R. (2016). Meaning matters: a short history of systemic functional linguistics. *Word* 61, 1–23. doi: 10.1080/00437956.2016.1141939
- Martin, J. R. (2017). Revisiting field: specialized knowledge in secondary school science and humanities discourse, *Onomázein* 2017, 111–148. [reprinted in *Accessing Academic Discourse: Systemic Functional Linguistics and Legitimation Code Theory*, eds J. R. Martin, K. Maton and Y. J. Doran. London: Routledge. 2020. 114–147].
- Martin, J. R. (2018). Interpersonal meaning: Systemic Functional Linguistics Perspectives. *J. Funct. Lang.* 25, 2–19. doi: 10.1075/fol.17018.mar
- Martin, J. R., Doran, Y. J., and Figueredo, G. (eds). (2020a). *Systemic Functional Language Description: Making Meaning Matter*. London: Routledge. doi: 10.4324/9781351184533
- Martin, J. R., and Doran, Y. J. (in press). Rethinking context: realisation, instantiation and individuation in SFL. *J. World Lang.*
- Martin, J. R., Maton, K., and Doran, Y. J. (2020b). “Academic discourse: an interdisciplinary dialogue,” in *Accessing Academic Discourse: Systemic Functional Linguistics and Legitimation Code Theory*, eds J. R. Martin, K. Maton, and Y. J. Doran (London: Routledge), 1–31. doi: 10.4324/9780429280726-1
- Martin, J. R., and Matruglio, E. (2013). “Revisiting mode: context in/dependency in ancient history classroom discourse,” in *Studies in Functional Linguistics and Discourse Analysis V*, eds G. Huang, D. Zhang, and X. Yang (Beijing: Higher Education Press), 72–95. [revised for Spanish translation by B Quiroz as ‘Retorno al modo: in/dependencia contextual en el discurso de las clases de historia antigua’. *Onomázein* (Número Especial IX ALSFAL) 2014, 186–213.] [English revision prepared for Spanish revision in *Accessing Academic Discourse: Systemic Functional Linguistics and Legitimation Code Theory*, eds J. R. Martin, K. Maton and Y. J. Doran. London: Routledge. 2020. 89–113] doi: 10.4324/9780429280726-4
- Martin, J. R., Quiroz, B., and Figueredo, G. (eds.). (2021). *Interpersonal Grammar: Systemic Functional Linguistic Theory and Description*. Cambridge: Cambridge University Press.
- Martin, J. R., Quiroz, B., and Wang, P. (2023). *Systemic Functional Grammar: A text-based description of English, Spanish and Chinese*. Cambridge: Cambridge University Press. doi: 10.1017/9781009284950
- Martin, J. R., and Rose, D. (2008) *Genre Relations: Mapping Culture*. London: Equinox.
- Martin, J. R., and Unsworth, L. (2024). *Reading Images for Knowledge Building: Analyzing Infographics in School Science*. London: Routledge (Routledge Studies in Multimodality). doi: 10.4324/9781003164586
- Martin, J. R., Wang, P., and Zhu, Y. (2013). *Systemic Functional Grammar: a Next Step into the Theory – Axial Relations*. Beijing: Beijing University Press.
- Maton, K. (2011). “Theories and things: the semantics of disciplinarity,” in *Disciplinarity: Functional Linguistic and Disciplinary Perspectives*, eds F. Christie, and K. Maton (London: Continuum), 62–86.
- Maton, K. (2014). *Knowledge and Knowers: Towards a Realist Sociology of Education*. London: Routledge.
- Maton, K. (2016). “Legitimation code theory: building knowledge about knowledge-building,” in *Knowledge-Building: Educational Studies in Legitimation Code Theory*, eds K. Maton, S. Hood, and S. Shay (London: Routledge), 1–24. doi: 10.4324/9781315672342
- Maton, K., and Chen, R. T. H. (2016). “LCT in qualitative research: creating a translation device for studying constructivist pedagogy,” in *Knowledge-Building: Educational Studies in Legitimation Code Theory*, eds K. Maton, S. Hood, and S. Shay (London: Routledge), 27–48.
- Maton, K., and Howard, S. K. (2016). “LCT in mixed-methods research: evolving an instrument for quantitative data,” in *Knowledge-Building: Educational Studies in Legitimation Code Theory*, eds K. Maton, S. Hood, and S. Shay (London: Routledge), 49–71.
- Maton, K., Martin, J. R., and Matruglio, E. (2016). “LCT and systemic functional linguistics: enacting complementary theories for explanatory power,” in *Knowledge-Building: Educational Studies in Legitimation Code Theory*, eds K. Maton, S. Hood, and S. Shay (London: Routledge), 93–114.
- Matthiessen, C. M. I. M. (2004). “Descriptive motifs and generalisations,” in *Language Typology: a Functional Perspective*, eds A. Caffarel, J. R. Martin, and C. M. I. M. Matthiessen (Amsterdam: Benjamins), 537–673. doi: 10.1075/cilt.253.12mat
- Moore, R., and Muller, J. (2002). The growth of knowledge and the discursive gap. *Br. J. Sociol. Educ.* 23, 627–637. doi: 10.1080/014256902000038477
- Muller, J. (2000). *Reclaiming Knowledge: Social Theory, Curriculum and Education Policy*. London: Routledge.
- Muller, J. (2007). “On splitting hairs: hierarchy, knowledge and the school curriculum,” in *Language, Knowledge and Pedagogy*, eds F. Christie and K. Maton (London: Continuum), 65–86.
- Muller, J. (2011). “Through others’ eyes: the fate of disciplines,” in *Disciplinarity: Functional Linguistic and Disciplinary Perspectives*, eds F. Christie, and K. Maton (London: Continuum), 13–34.
- Ngo, T., Hood, S., Martin, J. R., Painter, C., Smith, B. A., Zappavigna, M., et al. (2022). *Modelling Paralanguage Using Systemic Functional Semiotics: Theory and Application*. London: Bloomsbury. doi: 10.5040/9781350074934
- O’Toole, M. (1994). *The Language of Displayed Art*. London: Leicester University Press.
- Painter, C., Martin, J. R., and Unsworth, L. (2013). *Reading Visual Narratives: Image Analysis in Children’s Picture Books*. London: Equinox.
- Pike, K. L. (1988). “Bridging language learning, language analysis, and poetry, via experimental syntax,” in *Linguistics in Context: Connecting Observation and Understanding*, ed. D. Tannen (Norwood, NJ: Ablex), 221–245.
- Rose, D., and Martin, J. R. (2012). *Learning to Write, Reading to Learn: Genre, Knowledge and Pedagogy in the Sydney School*. London: Equinox.
- Saussure, F. (1916/1959). *Course in General Linguistics*. New York, NY: Philosophical Library.
- Thompson, G., Bowcher, W. L., Fontaine, L., and Schöenthal, D. (eds.). (2019) *The Cambridge Handbook of Systemic Functional Linguistics*. Cambridge: Cambridge University Press.
- Tseng, C., and Bateman, J. A. (2011). Multimodal narrative construction in Christopher Nolan’s *Memento*: a description of analytic method. *Vis. Commun.* 11, 91–119. doi: 10.1177/1470357211424691
- van Leeuwen, T. (1999). *Speech, Music, Sound*. Basingstoke: Macmillan. doi: 10.1007/978-1-349-27700-1
- van Leeuwen, T. (2005). “Three models of interdisciplinarity” in *New Agenda in (Critical) Discourse Analysis: Theory, Methodology and Interdisciplinarity*, eds R. Wodak, and P. Chilton (Amsterdam: Benjamins), 3–18. doi: 10.1075/dapsac.13.04lee
- van Leeuwen, T. (2011). *The Language of Colour*. London: Routledge.
- Whorf, B. L. (1945). Grammatical categories. *Language* 21, 1–11. [Reprinted in Carrol, J. B. (Ed.) (1956) *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf* (pp. 87–101). Cambridge, MA: The MIT Press.]
- Wignell, P. (2007a). “Vertical and horizontal discourse and the social sciences,” in *Language, Knowledge and Pedagogy*, eds F. Christie, and K. Maton (London: Continuum), 184–204.
- Wignell, P. (2007b). *On the Discourse of Social Science*. Darwin: Darwin University Press.
- Wildfeuer, J. (2021). “Discourse semantics and textual logic: methodological considerations for multimodal analysis.” In *OBST - Osnabrücker Beiträge zur Sprachtheorie 99. Special Issue on Linguistic and Multimodality*, eds A. Krause, and U. Schmitz, 87–114.
- Yu, Z. (forthcoming). *Multimodal Knowledge Building in Secondary School Chemistry Textbooks*. London: Bloomsbury.
- Zappavigna, M., and Logi, L. (2024). *Emoji and Social Media Paralanguages*. Cambridge: Cambridge University Press.





## OPEN ACCESS

## EDITED BY

Claudia Lehmann,  
University of Potsdam, Germany

## REVIEWED BY

Tiago Timponi Torrent,  
Federal University of Juiz de Fora, Brazil  
Alin Olteanu,  
RWTH Aachen University, Germany

## \*CORRESPONDENCE

Tuomo Hiippala  
✉ [tuomo.hiippala@helsinki.fi](mailto:tuomo.hiippala@helsinki.fi)

RECEIVED 13 November 2023

ACCEPTED 29 January 2024

PUBLISHED 12 February 2024

## CITATION

Hiippala T (2024) Rethinking multimodal corpora from the perspective of Peircean semiotics. *Front. Commun.* 9:1337434. doi: 10.3389/fcomm.2024.1337434

## COPYRIGHT

© 2024 Hiippala. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Rethinking multimodal corpora from the perspective of Peircean semiotics

Tuomo Hiippala\*

Department of Languages, University of Helsinki, Helsinki, Finland

This article discusses annotating and querying multimodal corpora from the perspective of Peircean semiotics. Corpora have had a significant impact on empirical research in the field of linguistics and are increasingly considered essential for multimodality research as well. I argue that Peircean semiotics can be used to gain a deeper understanding of multimodal corpora and rethink the way we work with them. I demonstrate the proposed approach in an empirical study, which uses Peircean semiotics to guide the process of querying multimodal corpora using computer vision and vector-based information retrieval. The results show that computer vision algorithms are restricted to particular domains of experience, which may be circumscribed using Peirce's theory of semiotics. However, the applicability of such algorithms may be extended using annotations, which capture aspects of meaning-making that remain beyond algorithms. Overall, the results suggest that the process of building and analysing multimodal corpora should be actively theorized in order to identify new ways of working with the information stored in them, particularly in terms of dividing the annotation tasks between humans and algorithms.

## KEYWORDS

multimodality, corpora, vector search, computer vision, Peirce, semiotics

## 1 Introduction

Multimodality research has been characterized as a form of “applied semiotics” due to its strong orientation to data, which distinguishes the field from mainstream semiotics (Bateman and Hiippala, 2021, p. 66). This orientation may be traced back—at least partially—to the influence of linguistics, which has a long history of studying language and its use from various perspectives and at various levels of abstraction. This kind of broad engagement with language required linguistics to develop robust methodologies for taking on diverse forms of linguistic data and phenomena. Not surprisingly, the field of linguistics was among the first to expand its research interests to considering how language and other modes of communication co-operate in making and exchanging meanings—a phenomenon now conceptualized as *multimodality*. Bateman (2022b) argues that such an extension beyond the traditional disciplinary borders reflects the nature of *multimodality as a stage of development* within a discipline, a process that can bring different disciplines concerned with similar data or phenomena into contact with each other. In addition to theories and frameworks, each discipline is likely to bring its own methodologies and ways of working with data to the contact situation.

Bearing this in mind, in this article I seek to problematise certain methodological imports from linguistics to multimodality research, focusing especially on methods for building and analysing multimodal corpora. The success of corpus methods, which form a major pillar of contemporary linguistics, may be ascribed to the availability of increasingly large volumes of annotated data and powerful methods for searching this data for patterns (Bateman, 2014, p. 239). Owing to their success in linguistics, corpus methods have been proposed as being useful for the field of multimodality research as well. Whereas some approaches to corpus-driven research on multimodality draw on corpus linguistic techniques for building annotation frameworks (see e.g. the use of stand-off annotations in Bateman, 2008), others advocate for a more direct application of linguistic corpus methods (see e.g. the use of concordancers in Christiansen et al., 2020). Instead of engaging with debates on which corpus linguistic techniques may be applicable to multimodality research and for what purposes, I take a step back and consider how we secure access to information stored in multimodal corpora more generally, and how this understanding may benefit their annotation and analysis. To do so, I approach the issue by drawing on the theory of semiotics developed by Charles Sanders Peirce (see e.g. Atkin, 2023), which has been previously brought into contact with theories of multimodality (Bateman, 2018) and multimodal corpora (Allwood, 2008).

## 2 Corpus-driven research on multimodality

Diverse research communities that study multimodality consider annotated corpora to be essential for conducting empirical research on the phenomenon (see e.g. Allwood, 2008; Bateman, 2014; Huang, 2021). In the context of multimodality research, an annotated corpus may be broadly defined as a collection of data about communicative situations or artifacts, which has been enriched with additional information about the data that is considered relevant for the research questions being asked. This kind of ‘data about data’ may range from generic *metadata* associated with individual entries in the corpus (such as information about author(s), date of publication, etc.) to multiple layers of cross-referenced annotations that allow combining information across annotation layers, which are needed for capturing the structure of multimodal discourse (see e.g. Bateman, 2008; Hiippala, 2015). These annotations, which are typically created using standardized markup languages such as XML or JSON, make working with the corpus tractable by allowing users to query the corpus for instances of particular annotations.

Corpus-driven empirical research has been viewed as crucial for establishing a stronger bond between theory and data in multimodality research (Bateman et al., 2004). Writing 20 years ago, Kaltenbacher (2004, p. 202) identified the lack of empiricism as a major weakness of the emerging field of study. Researchers working at that time sought to address this situation by developing annotation frameworks for multimodal corpora (Bateman et al., 2004) and linguistically-inspired concordancers for detecting patterns in transcripts of multimodal data (Baldry,

2004; Thomas, 2007) and identified challenges involved in applying corpus methods in multimodality research (Gu, 2006). More recently, Pflaeging et al. (2021, p. 3–4) have observed that empirical research on multimodality continues to be oriented toward qualitative research and small-scale studies using limited volumes of data (see also Bateman, 2022a). According to Pflaeging et al. (2021, p. 4), many multimodality researchers still hesitate to ‘scale up’ and increase the volume of data for various reasons: the work may simply be at a stage of development in which large-scale studies are not yet feasible, or there might be a lack of knowledge how to pursue such analyses altogether.

Although Pflaeging et al. (2021) discuss the nature of empirical multimodality research more generally, any efforts to scale up the volume of data are likely to involve the creation of annotated corpora, as annotations are needed for securing analytical access to the data in the corpus (Bateman et al., 2004, p. 69). However, large annotated corpora have remained elusive, because applying complex annotation frameworks to multimodal data requires time, resources and expertise. Hiippala et al. (2021), for example, present a corpus of 1,000 primary school science diagrams, which are annotated for their expressive resources, compositionality and discourse structure. The annotations were created over a period of six months by five research assistants trained to apply the annotation schema, which cost approximately 50,000€ (Hiippala et al., 2021, p. 673). Given the costs and resources needed for building corpora, it is not surprising that various proposals have been put forward for improving the efficiency of building multimodal corpora. These proposals range from using computational methods for automating parts of the annotation process (Bateman et al., 2016; Hiippala, 2016; O’Halloran et al., 2018; Steen et al., 2018) to paying crowdsourced non-expert workers available on online platforms to perform the annotation tasks (Hiippala et al., 2022).

Despite the recent advances, corpus methods and their application in multimodality research remain a long way from the level of methodological maturity achieved by corpus linguistics, which has established methods for data collection and annotating and querying corpora (see e.g. Lüdeling and Kytö, 2008). In this context, however, it should be noted that multimodality research seeks to apply corpus methods to data with diverse material properties and multiple semiotic modes. Whereas corpus linguistics could exploit the linear structure of spoken and written language for developing methods such as collocation analyses and keyword-in-context queries, multimodality research regularly takes on data whose materialities vary along the dimensions of temporality, space, participant roles and transience (Bateman, 2021). In terms of materiality, compiling a corpus that describes the multimodality of static, 2D page-based documents is radically different from building a corpus of communicative situations involving face-to-face interaction, which unfold in time and are construed dynamically by their participants. These material differences define to what extent a corpus may capture the multimodal characteristics of the artifacts or situations under analysis (Gu, 2006). In addition, this material diversity has implications for developing corpus methods for multimodality research, which must account for the properties of the underlying

materiality in order to make potentially meaning-bearing features accessible for analysis.

The importance of making the information stored in multimodal corpora accessible is emphasized by Bateman (2008, p. 251), who observes that:

... corpus-based research is all about searching for reoccurring patterns; the more the format of stored data can be made to support the activity of searching for patterns, then the more valuable that corpus becomes for analysis.

Arguably, the search for patterns may be supported by designing corpus annotation frameworks that adequately ‘expose’ the potentially meaning-carrying dimensions of materiality for annotation and analysis. In other words, the frameworks must inherently support annotating and retrieving information about semiotic modes that may be *potentially* deployed on the underlying materiality. To exemplify, an annotation framework targeting audiovisual media such as film, animation, television or video games must ensure that both temporal and spatial dimensions of the materiality are made available for description, as both may carry meaningful organizations of semiotic modes (see e.g. Stamenković and Wildfeuer, 2021). Along the temporal dimension, the framework must allow segmenting the data into shots, turns, actions or other basic temporal units, whose position along the timeline may be defined using timestamps. At each point in time, the framework must also allow decomposing the spatial dimension into analytical units, whose position in the layout space may be represented using coordinates. Finally, the framework must also allow synchronizing the descriptions across these temporal and spatial “canvases” (Bateman et al., 2017, p. 87) in order to account for their coordinated use for meaning-making. As Bateman et al. (2021, p. 116) note, it is entirely natural for multimodal artifacts to exhibit structures that unfold temporally and spatially, but their joint description is not necessarily supported by contemporary annotation software (see, however, Belcavello et al., 2022).

However, ensuring that the corpus design adequately exposes the material properties of the data is only a starting point for building multimodal corpora, as it provides a foundation for developing more sophisticated annotation frameworks that pick out characteristics of the semiotic modes deployed on these materialities. The functions of such annotations may range, for example, from identifying, categorizing and describing units of analysis to annotating their interrelations (see e.g. Bateman, 2008; Stöckl and Pflaeging, 2022). On a more general level, all annotation frameworks may be treated as semiotic constructs, whose complexity depends on the kinds of phenomena that the annotation framework seeks to capture. Given that these semiotic constructs are designed and reflect properties of the underlying data, I argue that the relationship between the annotations and the underlying data warrants additional attention, as this inevitably affects our ability to retrieve information from corpora.

### 3 A Peircean perspective to multimodal corpora

Compared to the efforts to build larger multimodal corpora, relatively little attention has been paid to how we are able to secure

any kind of access at all to the information stored in annotated corpora. One perspective to theorizing this issue may be provided by Peircean semiotics, which posits that access to “information” is mediated by signs and processes of signification (Bateman, 2018, p. 3). Allwood (2008, p. 209), who approaches multimodal corpora from a semiotic perspective, observes that multimodal corpora often feature signs that belong to three categories defined by Charles Sanders Peirce: icons, indices and symbols. He points out that static images, audiovisual moving images, sound recordings and many other forms of data stored in multimodal corpora are inherently *iconic*, because they bear resemblance to the original objects that they represent. According to Allwood (2008, p. 209), the iconic signs that make up a corpus may contain further indexical, iconic and symbolic signs—as exemplified by a sound recording (iconic) of human speech (symbolic). In addition, “raw” corpus data may be complemented by symbolic signs in the form of textual annotations, which can add “focus, identification and perspective” (Allwood, 2008, p. 209).

However, icons, indices and symbols cover only a part of Peirce’s theory of signs (Atkin, 2023). This is why considering multimodal corpora from a semiotic perspective may benefit from Bateman’s (2018) exploration of Peircean semiotics and its relationship to contemporary theories of multimodality. Bateman (2018, p. 3) emphasizes the phenomenological orientation of Peirce’s theory, which focuses on the human experience and attempts to capture “the nature of what could be known” (Jappy, 2013, p. 66). In Peirce’s view, signs do not reflect some pre-existing body of knowledge, but actively construe our lived experience. This orientation is evident in three categories proposed by Peirce—Firstness, Secondness and Thirdness—that provide a foundation for his theory of semiotics by carving out different ways of accessing information about the world (Bateman, 2018, p. 5). *Firstness* covers “independent” forms of signification, such as colors, shapes, textures and other qualities that are inherent to whatever is being interpreted. *Secondness* refers to forms of signification that pick out pairs of phenomena that depend on each other, as exemplified by the way smoke depends on fire. Finally, *Thirdness* stands for forms of signification based on conventional relations between entities, which can only be established by an external interpreter who construes a sign.

These categories are fundamental for understanding Peirce’s theory of semiotics, beginning with his definition of a sign. For Peirce, a sign involves three interrelated roles that need to be fulfilled in order to know more about something: if some role remains unfulfilled, there is no sign (Bateman, 2018, p. 6). First, the *sign-vehicle* (or representamen) stands for whatever that acts as the source of “information”. The sign-vehicle may range from a puff of smoke or the sound of a raindrop hitting the windowsill to an utterance, a drawing or a shape. Second, the *object* refers to the entity picked out by the sign-vehicle. The object also places constraints on the sign-vehicle, which the sign-vehicle must meet in order to be associated with the object (Bateman, 2018, p. 6). To exemplify, a sketch of a dog (a sign-vehicle) must have certain qualities associated with dogs to be recognized as such (an object). Finally, the *interpretant* refers to something that the sign-user construes about the object via the sign-vehicle, which may range from mental constructs to certain feelings and dispositions (Bateman et al., 2017, p. 57).



According to Peirce, all signs necessarily fall under the category of Thirdness, as “signs do not exist in the world, they are made by interpreters” (Bateman, 2018, p. 6). This does not mean, however, that the categories of Firstness and Secondness would be irrelevant, because they provide a foundation for Peirce’s first trichotomy of *qualisigns*, *sinsigns* and *legisigns*, which are concerned with the nature of the *sign-vehicle*. Qualisigns refer to inherent qualities associated with a sign-vehicle, as exemplified by color, shape, texture, etc., which fall under the category of Firstness. However, Bateman (2018, p. 7) emphasizes that according to Peirce, qualisigns cannot exist without something that actually carries these qualities, which invokes the category of Secondness: Peirce uses the term *sinsign* to describe sign-vehicles that carry such qualities. Finally, a *legisign* stands for a sign-vehicle that relies on an established convention and thus falls within the category of Thirdness. Legisigns, which operate in a ‘law-like’ manner, can generate replicas of themselves, which are instantiated as *sinsigns* (Jappy, 2013, p. 33).

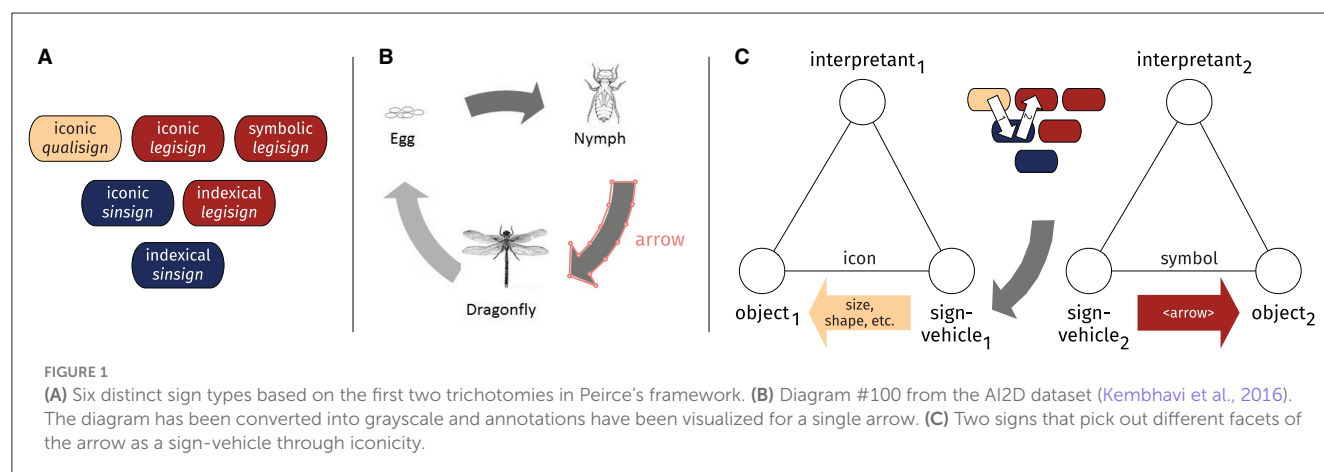
As pointed out above, the second trichotomy of *icons*, *indices* and *symbols* is arguably more widely known and used than the first trichotomy (cf. Allwood, 2008), although Peirce intended the trichotomies to be combined for describing how signs operate. They pick out different facets of semiosis, which need to be clearly demarcated, especially when applied to multimodal analysis (Bateman, 2018, p. 7–8). The second trichotomy is concerned with the relationship that holds between the sign-vehicle and the object (Bateman, 2018, p. 7). To begin with, icons are often understood as signs that rely on *resemblance based on shared properties*, although Bateman (2018, p. 4) argues that a more appropriate definition would involve treating iconicity as an “(abductive) hypothesis that a transferral of qualities makes sense”. Such hypotheses are by no means limited to visual properties. Indices, in turn, are commonly understood as being based on *causation*, that is, there exists a relationship between the object and the sign-vehicle, regardless whether this relation is established by some interpreter or not. In contrast to icons and indices, symbols rely on convention or agreement between sign users, which is why they constitute the least constrained form of signification (Bateman, 2018, p. 5).

Taken together, the first two trichotomies yield six distinct sign types, which are characterized by the kinds of “semiotic work” that they do (Bateman, 2018, p. 10). Here the internal

logic of the framework emerges from the relationships that hold between the categories of Thirdness, Secondness and Firstness, which describe the different ways gaining information about the world. Jappy (2013, p. 70) summarizes these interrelations as follows: any instance of Thirdness (a *legisign*) must be supported by Secondness through a *sinsign* that we recognize as a replica of the *legisign*. No *sinsign*, however, can be recognized as such without having particular qualities, which fall within the domain of Firstness, as they consist of *qualisigns*. In other words, Thirdness implies Secondness, which in turn implies Firstness (Jappy, 2013, p. 69–70). According to Bateman (2018, p. 10), these relationships can also be understood in terms of semiotic ‘power’, which defines just what kinds of “combinations of ways of being signs are licensed by the framework” proposed by Peirce.

The six sign types derived from the first two trichotomies are illustrated in Part A on the left-hand side of Figure 1 and colored according to their degree of semiotic “reach” in terms of Thirdness, Secondness and Firstness. As set out in Bateman (2018, p. 10), Thirdness (red) enables legisigns to be combined with icons, indices and symbols, whereas the Secondness (blue) of sinsigns limits them to icons and indices. Qualisigns (light brown), in turn, are limited to icons only. Note that these six sign types do not constitute a full account of Peirce’s sign types, as it lacks the sign types derived from the third trichotomy, which will be discussed shortly below in connection with the example shown in Figure 1B.

Figure 1B shows a single diagram from the AI2D dataset, which consists of nearly 5,000 primary school science diagrams that have been annotated for their features (Kembhavi et al., 2016). The diagram, which represents the life cycle of a dragonfly, has been converted from color to grayscale to highlight the annotations. For the purpose of exemplifying the annotations, a single bounding box that surrounds one of the arrows has been drawn on top of the original diagram image. The bounding box that traces the outline of the arrow consists of a polygon, which is essentially a series of coordinate points that indicates the location of the arrow in the diagram layout. This polygon is accompanied by the textual label ‘arrow’, which defines the type of the element designated by the polygon. Taken together, the polygon and the textual label represent common types of co-operating annotations found in multimodal corpora that are used to describe parts of the underlying artifact (see e.g. Bateman, 2008; Hiippala et al., 2021).



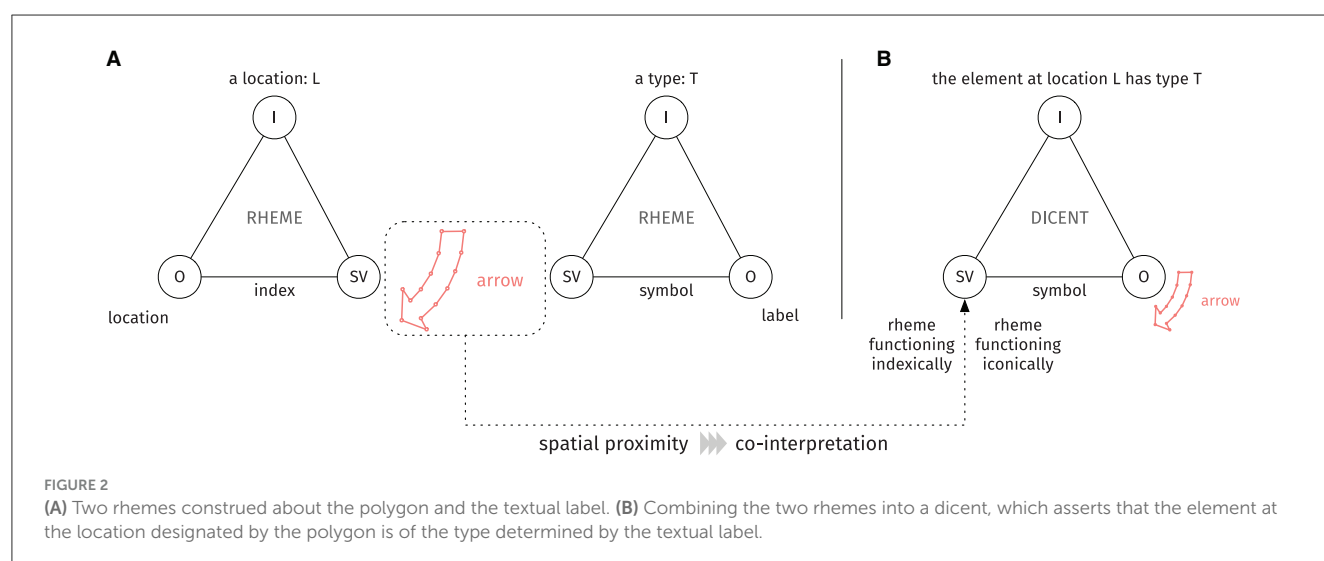
Having established the six sign types shown in Figure 1A, it is now possible to consider how they can be used to characterize aspects of the sign-making processes involved in annotating the diagram shown in Figure 1B. To do so, Figure 1C uses Peirce's tripartite model of a sign to represent two processes of signification that both pick out the element annotated as an arrow as the sign-vehicle. To begin with the sign on the left-hand side, if this element is taken as the sign-vehicle<sub>1</sub> of an *iconic qualisign*, the interpretant<sub>1</sub> construed about the element as an object<sub>1</sub> may concern, for example, its shape and texture. However, as pointed out above, qualities such as shape and texture fall within the domain of Firstness, and thus cannot exist independently. They must be inherent to some carrier, which in this case consists of the element as an *iconic sinsign* (not visualized in Figure 1C). Inferences made about the combined qualities of the sign-vehicle, such as its shape and texture, may lead to the conclusion that the element possesses qualities that are consistent with an arrow. The resulting interpretant may then entail properties inferred about the arrow, such as its direction and thickness.

Acknowledging that the arrow operates as an iconic sinsign can be used to push the analysis even further by considering its function in the diagram, as illustrated by the sign on the right-hand side of Figure 1C. If the annotator recognizes the arrow (sign-vehicle<sub>2</sub>) as a diagrammatic element (object<sub>2</sub>) that stands for a process (interpretant<sub>2</sub>), then this requires treating the arrow as a replica of an *iconic legisign* that governs the conventionalised use of arrows and lines for representing processes and relations in diagrams (see e.g. Alikhani and Stone, 2018; Lechner, 2020b). As noted above by Jappy (2013, p. 33), legisigns are replicated using sinsigns: they stand in a relationship of instantiation that Peirce described using the terms *type* and *token* (Bateman, 2018, p. 11). From a multimodal perspective, the conventionalised use of arrows, lines and other diagrammatic elements may be collectively characterized as an expressive resource commonly deployed within the diagrammatic semiotic mode (Hiippala and Bateman, 2022a). This also resonates with the proposal put forward in Bateman (2018, p. 20), who argues that semiotic modes may be conceptualized as specific kinds of legisigns that enable attributing meaning to forms deployed

on some materiality. Overall, the sign-making processes described above underline the continuous nature of semiosis (Jappy, 2013, p. 20) and how processes of signification enable a growth in knowledge (Bateman, 2018, p. 11), which is visualized in the middle of Figure 1C by movement from iconic qualisign to legisign via the sinsign.

The examples discussed above illustrate how Peircean semiotics can be used to describe the kinds of signs that may be construed about data stored in multimodal corpora. The same framework can be naturally applied to the annotations that describe the data as well. Returning to the arrow outlined in Figure 1B, the polygon stored in the corpus may be considered an indexical sinsign, which is a replica of a symbolic legisign. In this case, the symbolic legisign corresponds to a Cartesian coordinate system, which provides the mathematical and geometrical conventions needed for defining points in 2D layout space. The resulting sinsign may be treated as indexical, because its existence presumes that an annotator wanted to use the coordinate system as a symbolic legisign to demarcate a specific area of the diagram. This kind of motivated sign use is particularly important for annotations, which are assumed to reflect signs that the annotator has construed about the underlying data, as exemplified above by the arrow in Figure 1C. Some of these meanings may be captured using textual labels (Allwood, 2008, p. 209). In this case, associating the textual label 'arrow' with the polygon involves an indexical sinsign of another symbolic legisign, namely that of the English noun "arrow".

The way these two indexical sinsigns—the polygon and the textual label—co-operate in annotation can be described using the third Peircean trichotomy of *rhemes*, *dicents* and *arguments*. This trichotomy characterizes how the interpretant is shaped by the 'view' of the object provided by the sign-vehicle (Bateman, 2018, p. 12). To begin with, a rheme refers to something that may be construed about the sign-vehicle, such as a particular quality or a characteristic, but which cannot stand on its own due to its Firstness. To exemplify, the arrow in Figure 1B may be perceived as being wide or facing downward. Rhemes may be picked up in the second category of dicents, which combines rhemes into statements: one may assert, for example, that the arrow is wide



and faces down. Dicents are ‘independent’ and self-standing, and thus fall within the domain of Secondness. The final category of arguments comprises of multiple dicents combined into something that the individual construing the sign can take a stance on. One could construe an argument, for example, that arrows that are used to represent processes in primary school science diagrams tend to be wider than those used to pick out parts of some depicted object. This argument may be then accepted or rejected by the interpreter.

When viewed from the perspective of the third trichotomy, the polygon and the textual label “arrow” may be treated as *rhemes*, as they do not make assertions independently. Whereas the interpretant of the textual label defines the type of the element in question, the corresponding interpretant of the polygon determines its location, as visualized in Figure 2A. Due to their proximity in the diagram layout, the textual label and the polygon are likely to be interpreted together. Consequently, the two rhemes may be combined in the sign-vehicle of a dicent, which asserts that the element at the location designated by the polygon belongs to the category of arrows, as shown in Figure 2B. As Bateman (2018, p. 13) points out, Peirce considered the construction of dicents to be functionally constrained: one of the rhemes picked up as a part of the sign-vehicle must function indexically, whereas the other must function iconically. This ensures that any statement or assertion made by the dicent may be verified against the “evidence” provided. In this case, the polygon functions indexically by designating the location of the element, whereas the textual label functions iconically by positing that the qualities of the element marked out by the polygon are consistent with those of an arrow.

The example above illustrates how the annotations in multimodal corpora involve the co-operation of multiple sign types and shows how applying all three trichotomies can help sharpen the Peircean perspective to multimodal corpora offered in Allwood (2008). This arguably provides a deeper understanding of the semiotic underpinnings of corpus annotations, which can be used to rethink the way multimodal corpora are accessed and searched for patterns. Multimodal corpora generally rely on textual labels (rhemes) to describe the data, which are combined into dicents involving other rhemes, such as bounding boxes or timestamps. Emphasizing the role of textual labels as an access mechanism, Allwood (2008, p. 209) notes that “most existing multimodal corpora rely on textual identifying information in searching the corpus”, but Thomas (2014, p. 173) argues that “it is not always possible, nor is it necessarily productive, to describe every detail”. In particular, using textual labels to describe iconic qualities—such as size, color and shape—can prove challenging. Firstly, defining an exhaustive set of categories for systematically describing iconic qualities is likely to be difficult, and secondly, individual annotators may adopt different viewpoints to the data that are nevertheless equally valid (Gu, 2006, p. 129), which makes evaluating the reliability of the annotations difficult (see, however, Cabitza et al., 2023). This raises the question whether there are alternatives to using textual labels for accessing the information stored in multimodal corpora. As Allwood (2008) noted in 2008, “present technology mostly does not really allow efficient search using the iconic elements”, but this situation has now changed radically due to parallel work in the field of digital humanities, which has explored the

use of computational methods for detecting forms with similar qualities.

## 4 Computer vision in digital humanities and multimodality research

The rapidly expanding field of digital humanities now regularly engages with visual or multimodal materials, which often involves combining methods developed in the fields of computer vision, natural language processing and machine learning for enriching and exploring large volumes of data (Smits and Wevers, 2023). In addition to methodological explorations that have applied specific computational techniques to different media that range from film (Heftberger, 2018) to photography (Smits and Ros, 2023) and magic lantern slides (Smits and Kestemont, 2021) to mention just a few examples, recent research has sought to couch the application of computational methods to visual and multimodal materials within broader theoretical frameworks, such as the one proposed for “distant viewing” by Arnold and Tilton (2019, 2023). These efforts have also attracted the attention of multimodality researchers, who have argued that computational approaches to multimodal data in digital humanities would benefit from input from relevant theories of multimodality, which can provide the methodological tools needed for pulling apart the diverse materialities and artifacts studied (Bateman, 2017) and annotation schemes required for contextualizing the results of computational analyses (Hiippala, 2021).

On a trajectory parallel to digital humanities, there has been growing interest in the application of computational methods in multimodality research, but the use of these methods has been largely limited to annotating and analysing multimodal corpora. Hiippala and Bateman (2022b), for example, illustrate how combining computer vision and unsupervised machine learning allows describing the diversity of visual expressive resources (e.g. line drawings, colored illustrations) in the corpus of primary school science diagrams presented in Hiippala et al. (2021). Hiippala (2023), in turn, uses the same corpus to show how unsupervised machine learning can be used to identify diagram genres that are characterized by particular multimodal discourse patterns. Computational methods have also been used for automating parts of the annotation process for page-based (Hiippala, 2016) and audiovisual media (Bateman et al., 2016; Steen et al., 2018). O’Halloran et al. (2018), in turn, propose a mixed methods framework that combines qualitative multimodal analysis with quantitative techniques for data mining, whereas Thomas (2020) outlines strategies for applying computational methods in corpus-driven approaches to multimodality.

As pointed out above, much of the computational work in multimodality research is oriented toward analysing existing corpora or automating the creation of annotations. In contrast, many researchers working within the field of digital humanities have focused on developing methods for *retrieving* information from large collections of visual and multimodal data, which may not be accompanied by extensive metadata or annotations commonly expected of multimodal corpora (cf., however, Arnold

and Tilton, 2023). Here computer vision methods have proven especially useful, as they allow querying the data on the basis of formal properties such as texture, color and shape (Wasielewski, 2023, p. 40). One example of such an approach can be found in Lang and Ommer (2018), who show how computer vision methods can support iconographic research on visual arts, manuscripts and images. They present a system that allows the user to select an entire image or its part, which is then used for searching the dataset for visually similar occurrences. In other words, the ‘search term’ consists of an instance of data with particular iconic qualities.

The methods described above, which are now finding productive applications in fields such as digital art history (Wasielewski, 2023), have their roots in content-based image retrieval, a subfield of computer vision that develops methods for searching the content of images (see e.g. Smeulders et al., 2000). Because these methods are sufficiently generic to be applied in digital art history, it may be argued that they could also be applied to querying multimodal corpora, in which textual annotations remain the main way of securing access to the data. From a semiotic perspective, this would entail a major shift: instead of querying the data for instances of rhematic indexical sinsigns (e.g. specific instances of textual labels), the search criteria could be based on rhematic iconic qualisigns construed about the object of interest. This process is facilitated by rhematic indexical sinsigns in the form of bounding boxes (polygons or rectangles) that pick out parts of the underlying data. In other words, this would allow searching the corpora for instances of data that are similar in terms of visual qualities or *form*, as proposed by Lang and Ommer (2018). In the following sections, I explore the potential of such methods for multimodal corpus analysis by implementing a system that allows searching an existing corpus using iconic qualities.

## 5 Data and methods

The data of this study consists of two interrelated corpora. The first corpus, named AI2D-RST, contains 1,000 diagrams that represent topics in primary school natural sciences (Hiippala et al., 2021). The AI2D-RST corpus is a subset of the second corpus, the Allen Institute for Artificial Intelligence Diagrams dataset (AI2D; see Kembhavi et al., 2016). Whereas AI2D was developed for supporting research on automatic processing of diagrams, AI2D-RST is intended for studying diagrams as a mode of communication (Hiippala and Bateman, 2022a). AI2D contains crowdsourced non-expert annotations for diagram elements, their interrelations and position in diagram layout, which are loosely based on the work of Engelhardt (2002). The AI2D-RST corpus enhances the crowdsourced annotations provided in AI2D with expert annotations for compositionality, or how individual diagram elements are combined into larger units; discourse structure, or what kinds of relations hold between diagram elements; and connectivity, or how arrows and lines are used to set up connections between diagram elements or their groups.

Both AI2D and AI2D-RST use element types originally defined in AI2D: (1) text elements, (2) arrows, lines and other diagrammatic elements, (3) arrowheads and (4) blobs, which is a category that includes all forms of visual representation, such as illustrations,

line art, photographs, etc. (Hiippala et al., 2021, p. 665). In total, the 1,000 diagrams in AI2D-RST contain 20,094 elements categorized as text, arrows and blobs. I exclude arrowheads from the current analysis, as they simply augment the annotations for arrow elements. In addition to their type, each element is annotated for its position in the diagram layout. The coordinates for each element are represented using a polygon or a rectangle depending on the element type. The bounding boxes for blobs and arrows are represented using polygons, whereas text elements use rectangles, as illustrated in Figure 3. As such, the combinations of labels and bounding boxes constitute precisely the kinds of dicents described in Section 3 that allow retrieving information from the corpus.

I use the information about the position of each element in the diagram layout to extract them from the diagram image and describe their visual appearance using two computer vision algorithms, which approximate two iconic qualities: texture and shape. The first algorithm is *Local Binary Patterns* (LBP; Ojala et al., 1996), which is implemented in the *scikit-image* library for Python (van der Walt et al., 2014). The LBP algorithm describes the *texture* of an image. The operation of the algorithm and its applications in multimodality research have been described in Hiippala and Bateman (2022b, p. 418). More specifically, I use a rotation-invariant version of LBP, which means that the algorithm produces similar descriptions for images with similar textures regardless of their orientation. The output of the LBP algorithm consists of a 26-dimensional vector—a sequence of floating point numbers—that describes the texture of the image. The second algorithm is Zernike moments, which describes the *shape* of an image. Zernike moments are rotation- and scale-invariant, which means that they can capture similarities among shapes regardless of their size or orientation. I use the implementation of Zernike moments provided in the *mahotas* library for Python (Coelho, 2013), which yields a 25-dimensional vector that represents the shape of an image.

From a Peircean perspective, computer vision algorithms for low-level feature extraction, such as Local Binary Patterns for texture or Zernike moments for shape, are inherently constrained to the categories of Firstness and Secondness. Given some input data, the algorithms can seek to approximate qualities that fall within the domain of Firstness, which are then encoded into the sequence of numbers in the output vector. The resulting vector, whose existence and properties depend on the input data, may be considered a case of Secondness, because the vectors stand in an indexical relationship to the images they describe. These rhematic indexical sinsigns may be then use to model the iconic properties encoded within them (see Bateman, 2017, p. 37–38), but they are constrained to the domains of Firstness and Secondness (see Figure 1A). As Bateman (2018, p. 10) points out, one cannot “squeeze more semiotic ‘power’ out of a sign-situation than that sign-situation is configured to construe” due to the implication principle (Jappy, 2013, p. 69–70). In other words, the category of Thirdness remains beyond the reach of computer vision algorithms, as this would require an external interpreter for sign construction. This is extremely important to keep in mind when considering the capabilities of algorithms.

To store the output from the computer vision algorithms and to search for patterns, I use Milvus, an open-source vector



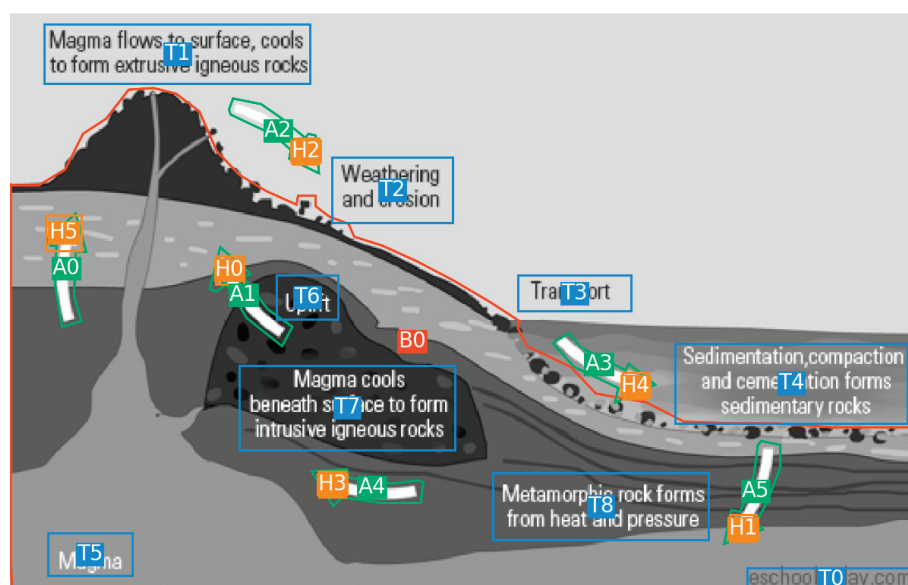


FIGURE 3

Diagram #4210 from the AI2D dataset. The diagram image has been converted to grayscale to highlight the crowdsourced annotations. The different elements picked out by the crowdsourced annotators are colored according to their type: text (blue), blobs (red), arrows (green) and arrowheads (orange). Each element is also accompanied by a unique identifier, e.g., T1 for the text element positioned in the upper left-hand corner. For a detailed analysis of this example and the AI2D annotation schema, see Hiippala and Bateman (2022a).

database for storing vectors and other data types, such as textual labels, Boolean values and integers (Wang et al., 2021). Milvus allows querying the database using a *vector search*, which involves defining a search vector that is then matched to other vectors in the database. For this purpose, Milvus implements various metrics for measuring the similarity of vectors, including Euclidean distance or cosine similarity. For current purposes, I use cosine similarity, which measures the similarity of vectors based on their direction and magnitude. The values for cosine similarity range from 1 for identical vectors to -1 for vectors that are exactly opposite in direction. A value of 0 indicates that the vectors are perpendicular, or at a 90° angle to each other. In addition to vector search, Milvus allows conducting a *hybrid search*, which searches for matches using both vectors and annotations stored in the database, such as textual labels that describe the type of element or diagram in question. For this reason, I enrich the entry for each diagram element in the database with information on diagram type from both AI2D and AI2D-RST (see Hiippala and Bateman, 2022b, p. 416) and the element type (text, arrow, blob).

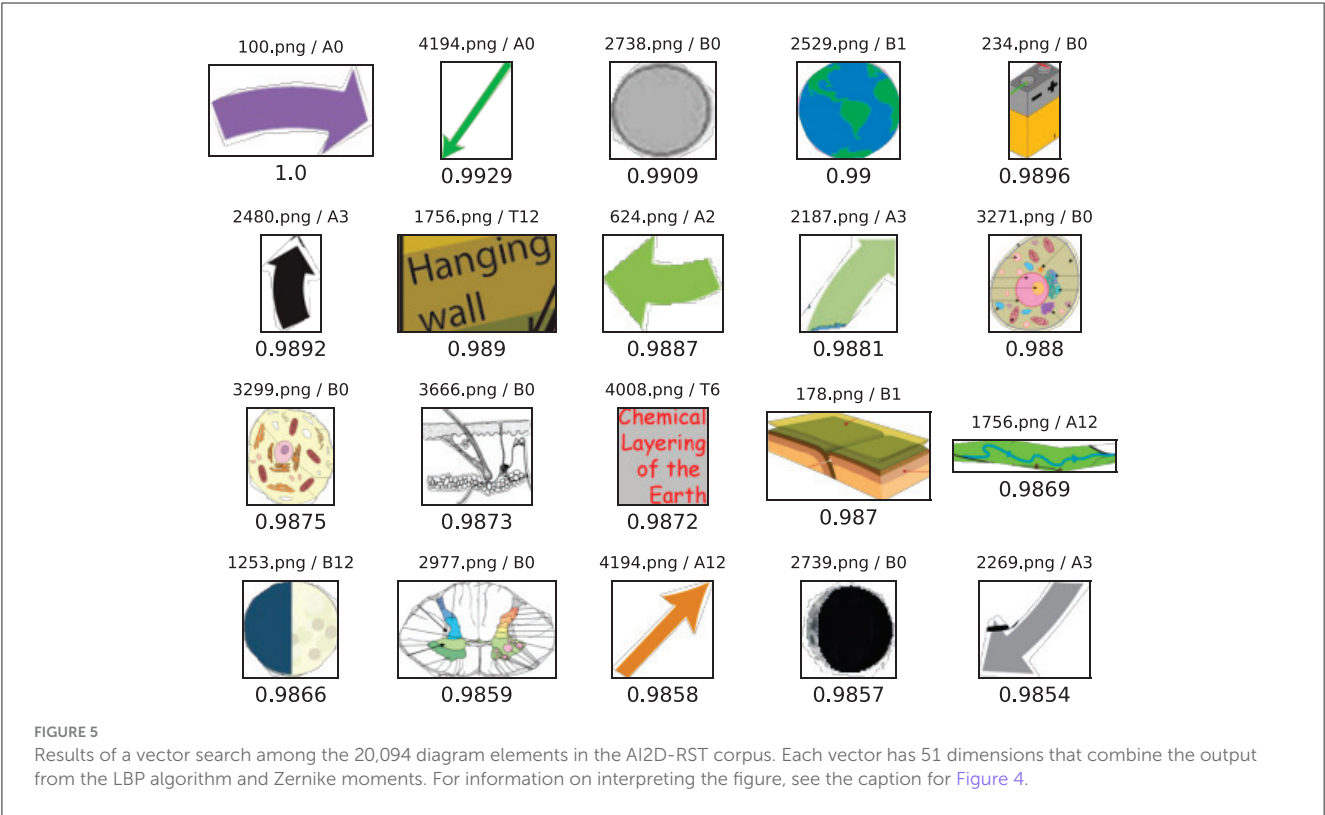
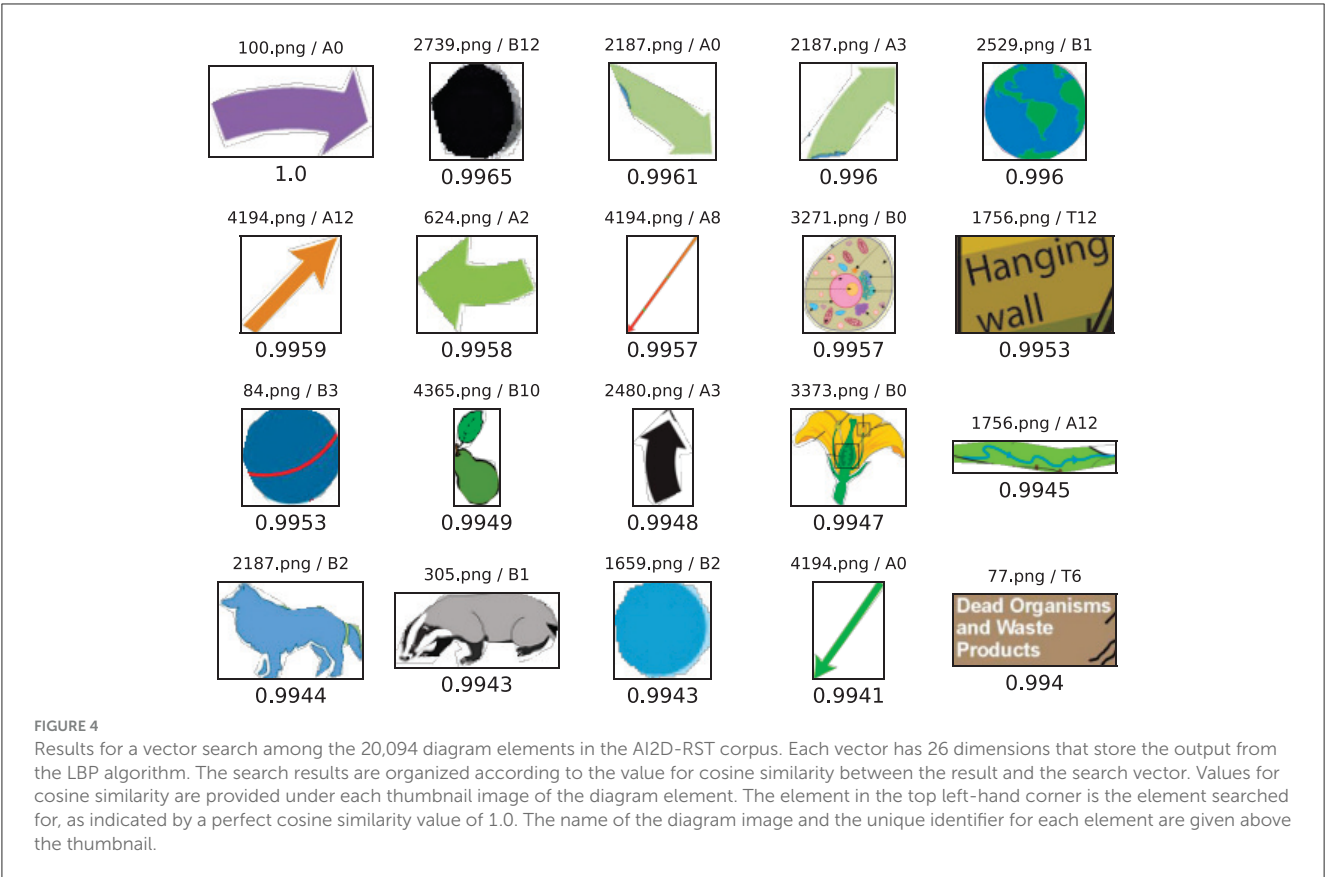
For reproducibility, the Python code for extracting data from the corpora, applying the computer vision algorithms and creating and querying the database is provided openly at: <https://doi.org/10.5281/zenodo.10566132>.

## 6 Analysis

To assess the potential of using computer vision and vector databases for querying multimodal corpora, I explore the use of arrows and lines as an expressive resource of the diagrammatic semiotic mode (Hiippala and Bateman, 2022a) in the AI2D-RST

corpus (Hiippala et al., 2021). Previous research has shown that diagrams regularly use arrows and lines for diverse communicative functions: they can, for example, represent processes and relationships that hold between diagram elements (see e.g. Alikhani and Stone, 2018; Lechner, 2020b). Lechner (2020a, p. 118), who explores how data visualizations use connecting lines to express uncertainty, observes that the iconic qualities of arrows and lines can determine or complement their communicative functions. She identifies various potentially meaning-bearing qualities of arrows and lines, such as orientation, size, color, pattern, etc., which can be used as the basis for annotating these properties (Lechner, 2020a, p. 117). However, as pointed out in Section 3, defining an annotation schema that seeks to capture iconic qualisigns construed about arrows and lines would likely require excessive time and resources due to the number of potentially meaningful qualities and raise questions about the reliability of the annotations (Thomas, 2014, p. 173). Given that the AI2D-RST corpus does not include annotations that describe the form or qualities of individual instances of expressive resources, but simply places them into abstract categories such as text, blobs and arrows, my aim is to evaluate whether the computational methods described in Section 5 can be used to retrieve visually similar arrows and lines from the AI2D-RST corpus, thus sidestepping the need to use textual labels for describing visual qualities.

Figure 4 shows the results of a vector search among the 20,094 elements categorized as text, arrows or blobs in the AI2D-RST corpus. Each element is processed using the LBP algorithm, which yields a 26-dimensional vector that describes the texture of the element. As explicated in Section 5, Milvus compares the search vector to each vector stored in the database and returns those that are closest to the search vector in terms of cosine similarity. In this



case, the element that is being searched for is a thick, colored arrow with a solid texture, which is shown in the upper left-hand corner of Figure 4 and has a cosine similarity value of 1.0, which indicates

perfect similarity. As Figure 4 shows, the search results include several arrows with similar textures, but also contain numerous instances of other expressive resources, such as illustrations and

written language. The diversity of the results reflects the limitations of texture, which represents only one quality that may be construed about arrows and approximated by algorithms. As the results show, texture as a quality is by no means exclusive to arrows as an expressive resource (see [Djonov and van Leeuwen, 2011](#)). This suggests that retrieving elements with specific qualities that may correspond to instances of particular expressive resources – such as arrows and lines – requires placing additional constraints on the search in terms of *form*.

To this end, [Figure 5](#) combines the 26-dimensional vector for LBP that describes the texture of an element with a 25-dimensional vector for Zernike moments, which describes its shape. This combination yields a 51-dimensional vector for each element, which jointly encodes information about both texture and shape. As the results of the query show, combining LBP and Zernike moments yields somewhat different search results than those shown in [Figure 4](#). Just like above, the results are not limited to arrows and lines, but also include instances of other expressive resources, which the computer vision algorithms perceive as having similar visual qualities. This illustrates a challenge that [Thomas \(2020, p. 84\)](#) discusses in relation to supporting empirical research on multimodality using computational methods, which involves moving beyond low-level “regularities of form” and toward higher levels of abstraction. [Thomas \(2020, p. 84\)](#) characterizes this transition from a Peircean perspective as a move from iconic qualisigns to iconic legisigns. As [Figure 5](#) shows, approximating just some iconic qualities that may be attributed to arrows, such as texture and shape, are not sufficient for identifying indexical sinsigns that could be potentially ascribed to the diagrammatic semiotic mode.

In light of the results shown in [Figures 4, 5](#), it should be emphasized that computer vision algorithms, such as LBP or Zernike moments, are inherently restricted to the domains of Firstness and Secondness, as pointed out in Section 5. Unlike humans, computer vision algorithms are not capable of the kind of continuous semiosis that enables the ‘growth’ of information (see [Figure 1C](#)). This kind of growth, which could entail a move to the domain of Thirdness, would be needed to recognize arrows as indexical sinsigns (replicas) generated by the diagrammatic semiotic mode ([Hiippala and Bateman, 2022a](#)). As a particular type of symbolic legisigns—highly conventionalised practices of manipulating materialities for communicative purposes that emerge within communities of users – semiotic modes fall within the domain of Thirdness ([Bateman, 2018, p. 20](#)). Because Thirdness remains beyond the reach of algorithms, mapping low-level regularities of form to more abstract categories needs to be supported by annotations, as noted by [Thomas \(2020, p. 84\)](#), in order to bridge what could be conceptualized as the “semiotic gap” (cf. [Smeulders et al., 2000](#)). From a Peircean perspective, the annotations needed for this purpose consist of textual labels and bounding boxes, which constitute rhemes that may be combined into a dicent that determines the type of the object at a given location (see [Figure 2](#)). In this case, a rhematic indexical sinsign—a replica of the English lexeme “arrow” as a symbolic legisign—provides sufficient “focus, identification and perspective” ([Allwood, 2008, p.](#)

209) for recovering indexical sinsigns that may be attributed to the diagrammatic mode.

In this way, the information provided by annotations enables shifting the direction of analysis from Thirdness toward Firstness (see [Bateman, 2018, p. 11](#)). Put differently, the annotations enable recovering information that remains unavailable to algorithms and limits their role to the domain of Firstness, that is, to describing iconic qualities. This approach may be implemented in a hybrid search, which combines a vector search with additional categorical or numerical information. [Figure 6](#) shows the results for a hybrid search, which compares the search vector consisting of LBP and Zernike moments to all other vectors in the database, but constrains the search to elements that have been annotated as arrows in the AI2D-RST corpus. As the results show, a hybrid search is able to retrieve arrows with similar iconic qualities in terms of texture and shape, regardless of their size or orientation. This is a notable result, as the application of these algorithms allows sidestepping the annotation of iconic qualities, which can consume excessive time and resources.

However, using a hybrid search to retrieve arrows with similar iconic qualities raises questions about quantifying the results, which is a common goal of pursuing corpus-driven analyses. Whereas annotations based on discrete labels can be counted and then subjected to statistical analyses, the results of a vector search are based on a continuous measure, in this case that of cosine similarity, which approximates their visual similarity. It is, however, possible to estimate the degree of similarity between the search vector and other vectors in the database. [Figure 7](#) plots the cosine similarity values between the search vector and (1) all arrows in the AI2D-RST corpus and (2) arrows in diagrams that have been categorized either as cycles or cut-outs ([Hiippala et al., 2021, p. 668](#)). As these plots show, the distributions of cosine similarity values do not enable visually identifying a cut-off point that could be used to determine which arrows are considered sufficiently similar to the one being searched for. However, potential differences in the distributions under different conditions can be evaluated statistically. In this case, a Mann-Whitney U-test indicates a statistically significant difference with a medium effect size between the samples for cycles and cut-outs ( $U = 351359, p = < 0.00$ , Cliff’s  $d = 0.369$ ), which suggests that these diagrams use arrows with different visual qualities.

When quantifying differences between iconic qualities with the help of computer vision and measures such as cosine similarity, one must naturally also consider the characteristics of the data in the corpus. Previous research has shown that cut-out diagrams are characterized by relatively stable layout patterns in which the depicted object is placed in the center of the layout, whereas the parts of the object are picked out using lines and written labels ([Hiippala and Bateman, 2022b; Hiippala, 2023](#)). Given that cut-out diagrams use lines to represent part-whole structures, it may be assumed that they would prefer to use thinner arrows and lines than cycles, which use these elements to represent processes and other phenomena ([Lechner, 2020b](#)). However, conducting a hybrid search for arrows among cut-out diagrams by using the same element as in [Figure 6](#) returns mixed results, which are shown in [Figure 8](#).



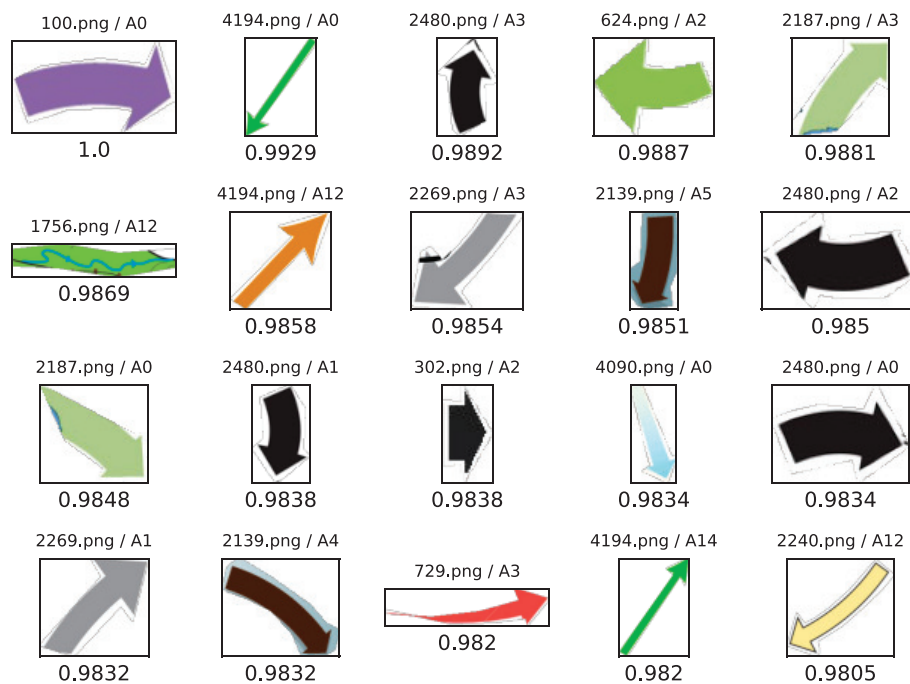


FIGURE 6

Results of a vector search among the 20,094 diagram elements in the AI2D-RST corpus. Each vector has 51 dimensions that combine the output from the LBP algorithm and Zernike moments. The output is constrained to elements that have been categorized as arrows in the AI2D-RST corpus. For information on how to interpret the results, see the caption for Figure 4.

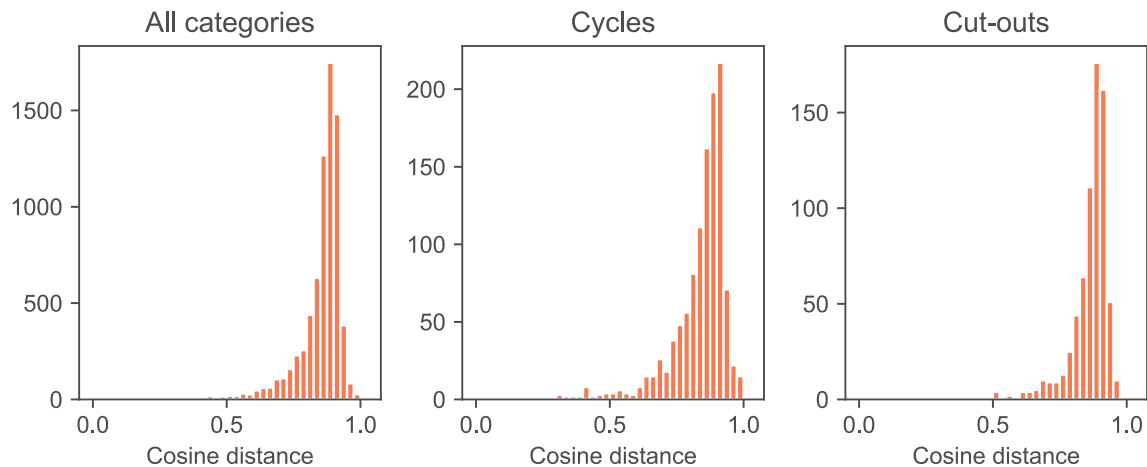


FIGURE 7

Cosine distances between the search vector and all other vectors in the database for all diagram types in the AI2D-RST dataset and cycle and cut-out diagrams (see Hiippala et al., 2021, p. 668).

The results show that cut-out diagrams do feature some wide arrows with solid texture, but many of the arrows returned by the vector search are indeed thinner, yet the algorithm considers them similar to the one that is being searched for. This may be traced back to inaccurate bounding boxes drawn by crowdsourced workers who annotated the data for the AI2D dataset (Kembhavi et al., 2016), which do not only include the arrow, but also cover parts of their immediate surroundings in the diagram. In other words, “thinness” is a quality that is difficult to capture using

polygons, but which also affects the results of a vector search. These surrounding areas may feature various shapes and textures, as illustrated by the examples in Figure 8. The gray area shows the extent of the bounding box: everything within the bounding box is provided as input to the computer vision algorithms, which results in “noise” that is encoded into the resulting vector representations. This also raises questions about the differences in the distribution of cosine distances in Figure 7, as capturing the property of thinness more accurately might make the differences between cut-outs and

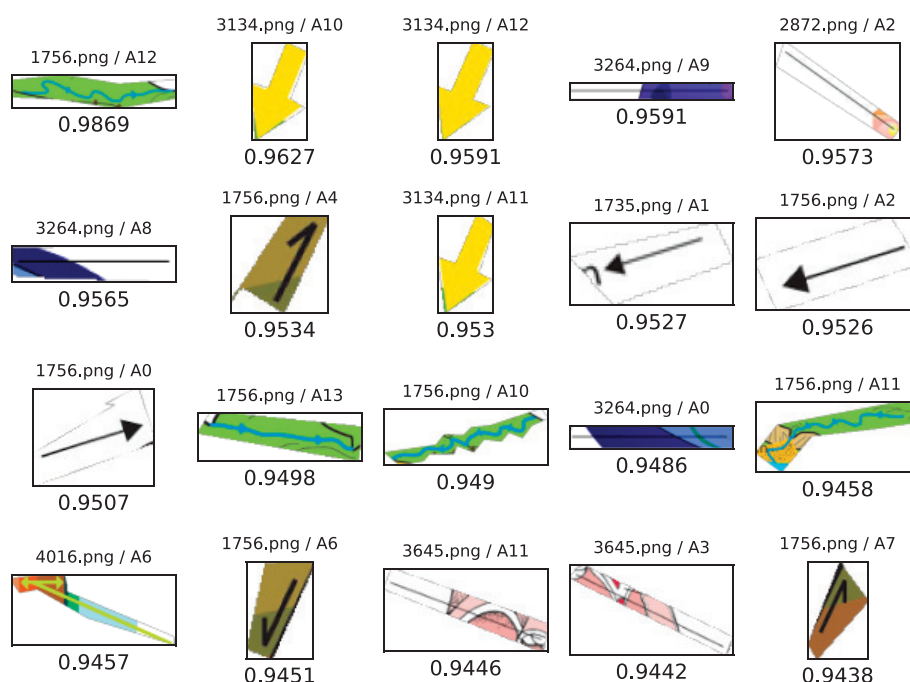


FIGURE 8

Results of a vector search among the 20,094 diagram elements in the AI2D-RST corpus. Each vector has 51 dimensions that combine the output from the LBP algorithm and Zernike moments. The output is constrained to elements that have been annotated as arrows in the AI2D-RST corpus, which occur in diagrams classified as cut-outs. For information on how to interpret the results, see the caption for Figure 4.

cycles more pronounced. To summarize, annotation quality has a significant impact on the applicability of computer vision and vector search methods for querying multimodal corpora.

## 7 Discussion

The results suggest that considering multimodal corpora from the perspective of Peircean semiotics benefits from a more comprehensive account that extends beyond the trichotomy of icons, indices and symbols (cf. Allwood, 2008). By providing a deeper understanding of corpus annotation frameworks as semiotic constructs that involve diverse types of signs, Peircean semiotics can be used to evaluate in what ways particular types of annotations are able to support access to the information stored in corpora. In particular, the results in Section 6 underline the importance of annotations as *dicent indexical sinsigns* that not only secure access to the data, but which can also constrain the operation of computer vision algorithms that operate on the elements designated by the annotations (see Figure 2). This information may be particularly useful for dividing the labor involved in annotating multimodal corpora. Although textual labels play a crucial role in securing access to the information stored in multimodal corpora, they may be less useful for describing iconic qualities, as many kinds of iconic qualisigns can be construed about the underlying data (Thomas, 2014, p. 173). This is precisely where computer vision methods may prove particularly useful.

From a Peircean perspective, the application of computer vision algorithms that approximate the qualities of forms present

on some materiality is necessarily constrained to the domain of Firstness. Whereas computer vision algorithms can approximate iconic qualities of the data and encode this information into numerical representations (a Second), the domain of Thirdness, which is a prerequisite for signification, remains beyond their reach. Nevertheless, the results show that computer vision algorithms can estimate iconic qualities of the underlying data when supported by annotations that constrain the search by providing information pertaining to the domain of Thirdness. Essentially, the annotations capture aspects of the signs that the human annotators have construed about the instances of data stored in the corpus. Although this process may be mimicked e.g., by training machine learning models to detect objects and predict labels associated with them, it should be noted that predicting a textual label for some entity—which is essentially a rhematic indexical sinsign—is an extremely constrained form of Secondness that does not enable the growth of information commonly attributed to semiosis, which may be considered a capability unique to humans. As such, the capabilities of such models with respect to processing visual and multimodal data should not be overestimated (cf. Arnold and Tilton, 2023). This also raises the question of how to collect high-level information pertaining to Thirdness at scale—using crowdsourced non-expert annotators available on crowdsourcing platforms presents one possible alternative (see Hiippala et al., 2022).

In terms of methodology, vector representations appear to hold much potential for supporting access to the information stored in multimodal corpora. As demonstrated in Section 6, hybrid searches may prove particularly useful, as they allow combining

low-level regularities of form captured by the vectors with higher-level information in the form of categorical labels, which has been identified as a key challenge in applying computational methods in multimodality research (Thomas, 2020, p. 84). It may also be argued that complementing traditional searches over annotated data with a vector search increases our capability to search multimodal corpora for patterns (Bateman, 2008, p. 251). However, whether a vector search is able to return results relevant to a query depends on the extent to which the algorithms used for creating the vectors are able to encode the properties of the data under analysis. This is especially important for multimodality research, as the search for patterns may not necessarily target particular kinds of objects, but rather attempts to retrieve instances of specific expressive resources such as written language, colored illustrations, line drawings, etc. As these expressive resources are characterized by particular forms, “traditional” computer vision algorithms based on human-designed heuristics, such as LBP or Zernike moments, may prove more useful than contemporary approaches involving deep neural networks (see e.g., Smits and Wevers, 2023), as these algorithms explicitly target formal properties such as shape or texture, and are less sensitive to rotation- and scale-invariance. It should also be noted that applying similar techniques to audiovisual data is likely to require different solutions (see e.g., Bateman et al., 2016).

For the design of multimodal corpora, the results suggest that additional attention should be paid to ensuring that the corpora support various means of access to the information stored therein, as this facilitates the search for patterns and thus makes the corpora more valuable for research (Bateman, 2008, p. 251). This means that rather than seeking maximum coverage in terms of annotation layers that rely on complex constellations of categories defined using textual labels, corpus design should carefully consider how different types of annotations interact with each other and what kind of information they provide access to. To exemplify, the results presented in Section 6 illustrate how polygons enable using computer vision to effectively extract information about the form of expressive resources directly from the corpus data, but this information only becomes usable when supported by textual labels that “add to, supplement and complement” the information made accessible by bounding boxes (Allwood, 2008, p. 209). Furthermore, the results underline that the quality of annotations remains of great importance: in addition to evaluating the reliability of analytical categories introduced by annotation schemas (Pflaeging et al., 2021, p. 21–22), one must ensure that the bounding boxes used to demarcate objects in the data are accurate, if computer vision methods are to be used for their analysis.

## 8 Conclusion

In this article, I have examined multimodal corpora from the perspective of Peircean semiotics. I have argued that Peircean semiotics can provide new perspectives on how multimodal corpora support access to the information stored in them. These perspectives are particularly valuable for designing, building and analysing multimodal corpora, as they help to determine what

kinds of descriptions are needed for capturing processes of meaning-making in communicative situations and artifacts. Given that creating multimodal corpora consumes excessive time and resources, Peircean semiotics can also be used to inform the division of labor between humans and computers. This kind of input from semiotics will be crucial as multimodal corpora begin to be extended to increasingly complex communicative situations and artifacts. This calls for increased efforts in theorizing the development and use of corpora in multimodality research, rather than considering corpus methods simply as a part of the methodological toolkit carried over from linguistics.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

TH: Conceptualization, Data curation, Formal analysis, Funding acquisition, Software, Visualization, Writing – original draft.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by a University of Helsinki three-year grant to the author.

## Acknowledgments

This article has benefitted considerably from discussions with John A. Bateman in the immediate past and over many years. Any errors or misrepresentations remain my responsibility.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

- Alikhani, M., and Stone, M. (2018). "Arrows are the verbs of diagrams," in *Proceedings of the 27th International Conference on Computational Linguistics* (Santa Fe, New Mexico: International Conference on Computational Linguistics), 3552–3563.
- Allwood, J. (2008). "Multimodal corpora," in *Corpus Linguistics: An International Handbook*, eds. A. Lüdeling and M. Kytö (Berlin: Mouton de Gruyter), 207–225.
- Arnold, T., and Tilton, L. (2019). Distant viewing: analyzing large visual corpora. *Digital Scholars. Human.* 34, i3–i16. doi: 10.1093/llc/fqz013
- Arnold, T., and Tilton, L. (2023). *Distant Viewing: Computational Exploration of Digital Images*. Cambridge, MA: MIT Press.
- Atkin, A. (2023). "Peirce's theory of signs," in *The Stanford Encyclopedia of Philosophy*, eds. E. N. Zalta and U. Nodelman (Stanford: Metaphysics Research Lab, Stanford University).
- Baldry, A. (2004). "Phase and transition, type and instance: patterns in media texts as seen through a multimodal concordancer," in *Multimodal Discourse Analysis: Systemic Functional Perspectives*, ed. K. L. O'Halloran (London: Continuum), 83–108.
- Bateman, J. A. (2008). *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*. London: Palgrave Macmillan.
- Bateman, J. A. (2014). "Using multimodal corpora for empirical research," in *The Routledge Handbook of Multimodal Analysis*, ed. C. Jewitt (London and New York: Routledge), second edn. 238–252.
- Bateman, J. A. (2018). Peircean semiotics and multimodality: towards a new synthesis. *Multimodal Commun.* 7:21. doi: 10.1515/mc-2017-0021
- Bateman, J. A. (2021). "Dimensions of materiality: towards an external language of description for empirical multimodality research," in *Empirical Multimodality Research: Methods, Evaluations, Implications*, eds. J. Pflaeging, J. Wildfeuer, and J. A. Bateman (Berlin and Boston: De Gruyter), 35–64.
- Bateman, J. A. (2022). Growing theory for practice: empirical multimodality beyond the case study. *Multimodal Commun.* 11, 63–74. doi: 10.1515/mc-2021-0006
- Bateman, J. A. (2022). Multimodality, where next? Some meta-methodological considerations. *Multimod. Soc.* 2, 41–63. doi: 10.1177/26349795211073043
- Bateman, J. A., Delin, J. L., and Henschel, R. (2004). "Multimodality and empiricism: preparing for a corpus-based approach to the study of multimodal meaning-making," in *Perspectives on Multimodality*, eds. E. Ventola, C. Charles, and M. Kaltenbacher (Amsterdam: Benjamins), 65–89.
- Bateman, J. A., and Hiippala, T. (2021). "From data to patterns: on the role of models in empirical multimodality research," in *Empirical Multimodality Research: Methods, Evaluations, Implications*, eds. J. Pflaeging, J. Wildfeuer, and J. A. Bateman (Berlin and Boston: De Gruyter), 65–90.
- Bateman, J. A., Thiele, L., and Akin, H. (2021). Explanation videos unravelled: breaking the waves. *J. Pragmat.* 175, 112–128. doi: 10.1016/j.pragma.2020.12.009
- Bateman, J. A., Tseng, C., Seizov, O., Jacobs, A., Lüdtkke, A., Müller, M. G., et al. (2016). Towards next generation visual archives: image, film and discourse. *Visual Stud.* 31, 131–154. doi: 10.1080/1472586X.2016.1173892
- Bateman, J. A., Wildfeuer, J., and Hiippala, T. (2017). *Multimodality: Foundations, Research and Analysis-A Problem-Oriented Introduction*. Berlin: De Gruyter Mouton.
- Bateman, J. A. (2017). Multimodale Semiotik und die theoretischen Grundlagen der Digital Humanities. *Zeitschrift für Semiotik* 39, 11–50.
- Belcavello, F., Viridiano, M., Matos, E., and Timponi Torrent, T. (2022). "Charon: A FrameNet annotation tool for multimodal corpora," in *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI)*, eds. S. Pradhan and S. Kuebler (Marseille, France: European Language Resources Association), 91–96.
- Cabitz, F., Campagner, A., and Basile, V. (2023). Toward a perspectivist turn in ground truthing for predictive computing. *Proc. AAAI Conf. Artif. Intell.* 37, 6860–6868. doi: 10.1609/aaai.v37i6.25840
- Christiansen, A., Dance, W., and Wild, A. (2020). "Constructing corpora from images and text: an introduction to visual constituent analysis," in *Corpus Approaches to Social Media*, eds. S. Rüdiger and D. Dayter (Amsterdam: Benjamins), 149–174.
- Coelho, L. P. (2013). Mahotas: Open source software for scriptable computer vision. *J. Open Res. Softw.* 1:e3. doi: 10.5334/jors.ac
- Djonov, E. N., and van Leeuwen, T. (2011). The semiotics of texture: from tactile to visual. *Visual Commun.* 10, 541–564. doi: 10.1177/1470357211415786
- Engelhardt, Y. (2002). *The Language of Graphics: A Framework for the Analysis of Syntax and Meaning in Maps, Charts and Diagrams* (Ph.D. thesis) Amsterdam: Institute for Logic, Language and Computation, University of Amsterdam.
- Gu, Y. (2006). Multimodal text analysis: a corpus linguistic approach to situated discourse. *Text & Talk* 26, 127–167. doi: 10.1515/TEXT.2006.007
- Heftberger, A. (2018). *Digital Humanities and Film Studies: Visualising Dziga Vertov's Work*. Cham: Springer.
- Hiippala, T. (2015). *The Structure of Multimodal Documents: An Empirical Approach*. New York and London: Routledge.
- Hiippala, T. (2016). "Semi-automated annotation of page-based documents within the Genre and Multimodality framework," in *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (Berlin, Germany: Association for Computational Linguistics), 84–89.
- Hiippala, T. (2021). Distant viewing and multimodality research: prospects and challenges. *Multimod. Soc.* 1, 134–152. doi: 10.1177/26349795211007094
- Hiippala, T., Alikhani, M., Haverinen, J., Kalliokoski, T., Logacheva, E., Orekhova, S., et al. (2021). AI2D-RST: a multimodal corpus of 1000 primary school science diagrams. *Lang. Resour. Evaluat.* 55, 661–688. doi: 10.1007/s10579-020-09517-1
- Hiippala, T., and Bateman, J. A. (2022). "Introducing the diagrammatic semiotic mode," in *Diagrammatic Representation and Inference: 13th International Conference (Diagrams 2022)*, eds. V. Giardino, S. Linker, R. Burns, F. Bellucci, J.-M. Boucheix, and P. Viana (Cham: Springer), 3–19.
- Hiippala, T., and Bateman, J. A. (2022). Semiotically-grounded distant view of diagrams: insights from two multimodal corpora. *Digit. Scholars. Human.* 37, 405–425. doi: 10.1093/llc/fqab063
- Hiippala, T., Hotti, H., and Suviranta, R. (2022). "Developing a tool for fair and reproducible use of paid crowdsourcing in the digital humanities," in *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (Gyeongju, Republic of Korea: International Conference on Computational Linguistics), 7–12.
- Hiippala, T. (2023). Corpus-based insights into multimodality and genre in primary school science diagrams. *Visual Commun.* doi: 10.1177/14703572231161829
- Huang, L. (2021). Toward multimodal corpus pragmatics: Rationale, case, and agenda. *Digit. Scholars. Human.* 36, 101–114. doi: 10.1093/llc/fqz080
- Jappy, T. (2013). *Introduction to Peircean Visual Semiotics*. London and New York: Bloomsbury.
- Kaltenbacher, M. (2004). Perspectives on multimodality: from the early beginnings to the state of the art. *Inform. Design J.* 12, 190–207. doi: 10.1075/ijdd.12.3.05kal
- Kembhavi, A., Salvato, M., Kolve, E., Seo, M. J., Hajishirzi, H., and Farhadi, A. (2016). "A diagram is worth a dozen images," in *Proceedings of the 14th European Conference on Computer Vision (ECCV 2016)* (Cham: Springer), 235–251.
- Lang, S., and Ommer, B. (2018). Attesting similarity: supporting the organization and study of art image collections with computer vision. *Digit. Scholars. Human.* 33, 845–856. doi: 10.1093/llc/fqy006
- Lechner, V. E. (2020a). "Modality and uncertainty in data visualizations: a corpus approach to the use of connecting lines," in *Diagrammatic Representation and Inference: 11th International Conference (Diagrams 2020)*, eds. A.-V. Pietarinen, P. Chapman, L. B. de Smet, V. Giardino, J. Corter, and S. Linker (Cham: Springer), 110–127.
- Lechner, V. E. (2020b). "What a line can say: Investigating the semiotic potential of the connecting line in data visualizations," in *Data Visualization in Society*, eds. H. Kennedy and M. Engebretsen (Amsterdam: Amsterdam University Press), 329–346.
- Lüdeling, A. and Kytö, M. (2008). *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter.
- O'Halloran, K. L., Tan, S., Pham, D.-S., Bateman, J. A., and Vande Moere, A. (2018). A digital mixed methods research design: integrating multimodal analysis with data mining and information visualization for big data analytics. *J. Mixed Methods Res.* 12, 11–30. doi: 10.1177/1558689816651015
- Ojala, T., Pietikäinen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition* 29, 51–59. doi: 10.1016/0031-3203(95)00067-4
- Pflaeging, J., Bateman, J. A., and Wildfeuer, J. (2021). "Empirical multimodality research: the state of play," in *Empirical Multimodality Research: Methods, Applications, Implications*, eds. J. Pflaeging, J. Wildfeuer, and J. A. Bateman (Berlin and Boston: De Gruyter), 3–32.
- Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 1349–1380. doi: 10.1109/34.895972
- Smits, T., and Kestemont, M. (2021). "Towards multimodal computational humanities: using CLIP to analyze late-nineteenth century magic lantern slides," in *Proceedings of the Computational Humanities Research Conference (CHR 2021)*, 149–158.
- Smits, T., and Ros, R. (2023). Distant reading 940,000 online circulations of 26 iconic photographs. *New Media Soc.* 25, 3543–3572. doi: 10.1177/1461448211049459
- Smits, T., and Wevers, M. (2023). A multimodal turn in digital humanities: Using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections. *Digit. Scholars. Human.* 38, 1267–1280. doi: 10.1093/llc/fqad008

- Stamenković, D., and Wildfeuer, J. (2021). "An empirical multimodal approach to open-world video games: a case study of Grand Theft Auto V," in *Empirical Multimodality Research: Methods, Evaluations, Implications*, eds. J. Pflaeging, J. Wildfeuer, and J. A. Bateman (Berlin and Boston: De Gruyter), 259–279.
- Steen, F. F., Hougaard, A., Joo, J., Olza, I., Cánovas, C. P., Pleshakova, A., et al. (2018). Toward an infrastructure for data-driven multimodal communication research. *Linguistics Vanguard* 4. doi: 10.1515/lingvan-2017-0041
- Stöckl, H., and Pflaeging, J. (2022). Multimodal coherence revisited: notes on the move from theory to data in annotating print advertisements. *Front. Commun.* 7. doi: 10.3389/fcomm.2022.900994
- Thomas, M. (2007). "Querying multimodal annotation: a concordancer for GeM," in *Proceedings of the Linguistic Annotation Workshop (LAW 2007)* (Prague, Czech Republic: Association for Computational Linguistics), 57–60.
- Thomas, M. (2014). Evidence and circularity in multimodal discourse analysis. *Visual Commun.* 13, 163–189. doi: 10.1177/1470357213516725
- Thomas, M. (2020). "Making a virtue of material values: tactical and strategic benefits for scaling multimodal analysis," in *Multimodality: Disciplinary Thoughts and the Challenge of Diversity*, eds. J. Wildfeuer, J. Pflaeging, J. A. Bateman, O. Seizov, and C. Tseng (Berlin: De Gruyter), 69–91.
- van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., et al. (2014). scikit-image: image processing in Python. *PeerJ* 2, 453. doi: 10.7717/peerj.453
- Wang, J., Yi, X., Guo, R., Jin, H., Xu, P., Li, S., et al. (2021). "Milvus: a purpose-built vector data management system," in *Proceedings of the 2021 International Conference on Management of Data* (New York, NY, USA: Association for Computing Machinery), 2614–2627.
- Wasielewski, A. (2023). *Computational Formalism: Art History and Machine Learning* (Cambridge, MA: MIT Press).





## OPEN ACCESS

## EDITED BY

Maria Grazia Sindoni,  
University of Messina, Italy

## REVIEWED BY

Kyle Jasmin,  
University of London, United Kingdom  
Valentin Werner,  
University of Bamberg, Germany

## \*CORRESPONDENCE

Claudia Lehmann

✉ claudia.lehmann@uni-potsdam.de

RECEIVED 15 November 2023

ACCEPTED 01 February 2024

PUBLISHED 19 February 2024

## CITATION

Lehmann C (2024) What makes a multimodal construction? Evidence for a prosodic mode in spoken English.

*Front. Commun.* 9:1338844.

doi: 10.3389/fcomm.2024.1338844

## COPYRIGHT

© 2024 Lehmann. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# What makes a multimodal construction? Evidence for a prosodic mode in spoken English

Claudia Lehmann\*

Chair of Present-Day English, Institute of English and American Studies, University of Potsdam, Potsdam, Germany

Traditionally, grammar deals with morphosyntax, and so does Construction Grammar. Prosody, in contrast, is deemed *paralinguistic*. Testifying to the “multimodal turn,” the past decade has witnessed a rise in interest in multimodal Construction Grammar, i.e., an interest in grammatic constructions other than exclusively morphosyntactic ones. Part of the debate in this recent area of interest is the question of what defines a multimodal construction and, more specifically, which role prosody plays. This paper will show that morphosyntax and prosody are two different semiotic modes and, therefore, can combine to form a multimodal construction. To this end, studies showing the independence of prosody for meaning-making will be reviewed and a small-scale experimental study on the ambiguous utterance *Tell me about it* will be reported on.

## KEYWORDS

Construction Grammar, usage-based, prosody, semiotic mode, forced-choice experiment

## 1 Introduction

Grammar deals with morphosyntactic patterns. True to this claim, the introductory sentence to the *Oxford Handbook of English Grammar* states that “grammar” is used in the sense which encompasses morphology (the principles of word formation) and syntax (the system for combining words into phrases, clauses, and sentences)” (Aarts et al., 2019). Construction Grammar is no exception to this rule: Goldberg defines a grammatical construction as a “learned pairing of form with semantic or discourse function, including morphemes or words, idioms, partially lexically filled and fully general phrasal patterns” (Goldberg, 2006, p. 5). While Construction Grammar foregrounds the role meaning plays in forming grammatical structures, neither intonation nor prosody are explicitly mentioned. This is surprising to the extent that research at the prosody-meaning interface has a long tradition and intonation is acknowledged to fulfill grammatical functions (see e.g., Tench, 1996; Wells, 2006; Levis and Wichmann, 2015; Nolan, 2021). One of the reasons for separating prosody from grammar may have to do with the fact that even within prosody research, its grammatical function used to be downplayed, maintaining that “in practice it is usually context that disambiguates and the role of intonation is minimal” (Levis and Wichmann, 2015, p. 151), even though Wichmann and Blakemore (2006, p. 1,537) argue earlier that “[t]he choice of a rise or fall, or the placement of a pitch accent, may be as important a cue to speaker meaning as its phonetic realization.” Rather, the so-called paralinguistic functions of prosody were foregrounded, i.e., its role in indicating emotions and attitudes (Féry, 2017, p. 7) and, indeed, the grammatical and the attitudinal functions of prosody are often interrelated (Gussenhoven, 2004).

Testifying to the “multimodal turn,” the past decade has witnessed a rise in interest in multimodal Construction Grammar (see Section 2.2 below), i.e., an interest in constructions other than exclusively morphosyntactic ones. Part of the debate in this recent area of interest is the question of what defines a multimodal construction and, more specifically, which role prosody plays. While it seems uncontested that the combination of a morphosyntactic and a kinesic form might form a multimodal construction (see e.g., Ningelgen and Auer, 2017; Ziem, 2017; and other papers in Zima and Bergs, 2017; or in Uhrig, 2020), prosodic peculiarities of constructions are seldom addressed (notable exceptions include Lelandais and Ferré, 2019; Pöldvere and Paradis, 2020). There is no *a priori* reason to exclude prosody from a constructional analysis, though; the only reason to do so seems to be the traditional misconception of prosody being something outside of the scope of grammar and, therefore, not worth any further consideration.

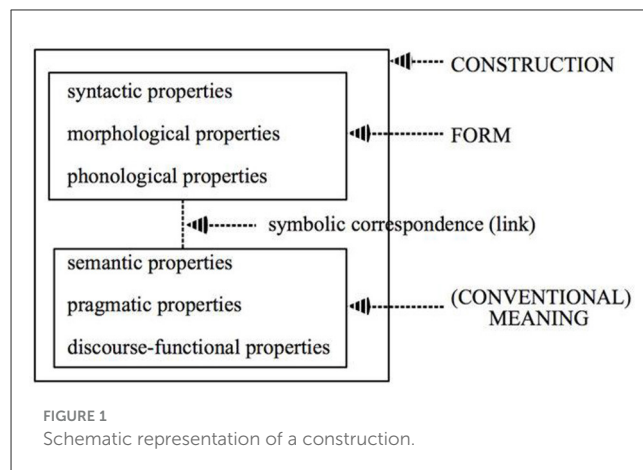
The aim of the present paper is twofold. First, it will show that prosody and morphosyntax can (and should) be considered independent semiotic modes (in the sense of Bateman et al., 2017), which independently can fulfill grammatical functions. Second, the paper will also show that the two semiotic modes can combine to form a multimodal construction (in the sense of Construction Grammar). The paper will proceed as follows: The main tenets of usage-based Construction Grammar and the notion of multimodal constructions will be introduced. Based on previous research, the paper will then argue that prosody and morphosyntax are independent semiotic modes by showing that they make use of different materiality and forms and that they independently contribute to the discourse semantics. It will then report on evidence that the two different modes may combine to form a multimodal construction using the results of a forced choice experiment.

## 2 (Usage-based) Construction Grammar and multimodality

In this section, the core assumptions of (usage-based) Construction Grammar and its relation to multimodality will be introduced. More specifically, the debate surrounding the notion of multimodal construction will be reviewed.

### 2.1 Constructions in Construction Grammar

Construction Grammar is no unified theory. For an overview of the different strands of Construction Grammar, Hoffmann and Trousdale (2013) is a useful resource. One of a few things all Construction Grammars have in common is that they consider the construction to be the core unit of language-related knowledge. A unit is considered a construction (C) “iff<sub>def</sub> C is a form-meaning pair  $\langle F_i, S_i \rangle$  such that some aspects of  $F_i$  or some aspect of  $S_i$  is not strictly predictable from C’s component parts or from other previously established constructions” (Goldberg, 1995, p. 4). Figure 1 provides a schematic representation of a



construction (taken from Croft and Cruse, 2004, p. 258). An example is the English idiom *Tell me about it*. Its component parts suggest (predict) that information is requested, but experienced language users know that it can also mean “I’m well aware of that; ‘I agree;’ ‘you don’t have to tell me’” (Tell, 2023). Since its meaning cannot be predicted from its component parts, it is a separate construction and must be learned. From such a perspective, idioms enjoy the same ontological status as words and more schematic constructions.

Usage-based approaches to Construction Grammar also consider predictable units to be constructions as long as they occur frequently enough so that they become entrenched in the language users constructicon, i.e., the mental repository of constructions (e.g., Bybee, 2006, 2013; Goldberg, 2006; Divjak, 2019). One example for this is the word *singer*. Even though its meaning “someone who sings” is perfectly predictable from its component parts, the verb *sing* and the derivational morpheme *-er*, the derivate *singer* is likely stored as a separate construction, because it is one of the 5,000 most frequent words in (written) English (Singer, 2023). Usage-based approaches to Construction Grammar further assume that the cognitive processes involved in language production and comprehension are domain-general and not specific to language. One of these domain-general cognitive processes is cross-modal association, which “allows humans to match up the phonetic (or manual) form experienced with properties of the context and meaning” (Bybee, 2013, p. 50), and which seems to be key in language learning (Imai and Kita, 2014; Dingemans et al., 2015). An example of cross-modal association is sound symbolism, which is more pervasive in English than traditionally assumed. Sidhu et al. (2021) could show that sounds associated with roundedness (like /m/) more often than not denote round objects in English, while sounds associated with spikiness (like /k/) often denote spiky objects in English; an effect also known as the maluma/takete effect (Köhler, 1929).

### 2.2 Multimodal constructions

Constructions can be of any size, “including morphemes or words, idioms, partially lexically filled and fully general

phrasal patterns” (Goldberg, 2006, p. 5) as well as argument and information structure constructions (see e.g., relevant chapters in Hoffmann and Trousdale, 2013; Hilpert, 2019; Hoffmann, 2022), but, evidently, the vast majority of constructions considered is of a morphosyntactic nature. This is surprising to the extent that usage-based Construction Grammar emphasizes language knowledge to emerge from the input language users get—and arguably this input commonly is multimodal. For instance, spoken language, i.e., the language infants are exposed to first, is inherently multimodal (Vigliocco et al., 2014; Feyaerts et al., 2017; Perniss, 2018), since speakers use gaze, gestures, facial expressions and other resources to convey meaning (see also Section 4.1 on the multimodality of *Tell me about it*). But also written language is often produced in multimodal situations (see e.g., Kress, 2000; van Leeuwen, 2014; Hiippala, 2017). Internet memes, for example, use written language and an image to convey their (conventionalized) meaning (Dancygier and Vandelanotte, 2017; Bülow et al., 2018). Despite these facts, multimodal constructional analyses are often noticeably absent from research in (usage-based) Construction Grammar.

In parallel to the multimodal turn in linguistics in general (see Stöckl, 2020), the past decade has also witnessed a growing interest in multimodal issues in Construction Grammar. One strand of research concerns itself with speech-embedded non-verbal depictions, i.e., gestures that may fill specific slots of constructions, such as Verb or Noun Phrase (see e.g., Clark, 2016; Ladewig, 2020; Hsu et al., 2021). Although not all of these studies position themselves in a Construction Grammar framework, their examples can be reanalyzed, like in Example (1):

- (1) [MB was discussing a measure in a Mozart sonata] But then he writes “(gazing at audience and singing) *dee-duh dum*.” That is very expressive.  
(Clark, 2016, p. 325)

From a Construction Grammar perspective, the nonverbal depiction (i.e., *dee-duh dum*) fulfills the function of the object noun phrase in the transitive construction. Examples like these thus show that constructional slots need not be filled by morphosyntactic elements but can also be realized by other means.

Another strand of research discusses the possible existence of multimodal constructions. Ziem (2017) names four conditions under which a construction can be seen as multimodal, of which only the first two will be reviewed here, because they are central to the argumentation put forward in this paper.<sup>1</sup> The first condition states that

- (a) A multimodal construction is a conventionalized pairing of a complex form that consists, at least, of a verbal element combined with a kinetic element (Ziem, 2017, p. 5).

In other words, a multimodal construction needs some kind of verbal form (with syntactic, morphological and/or phonological properties) and, necessarily, a kinetic element (like a manual gesture, a facial expression, or a particular gaze behavior) to be called such. Based on the representation of a construction (provided in Croft and Cruse, 2004, p. 258), Figure 2 depicts the representation of a multimodal construction.

A prime example for such a multimodal construction is the complex form of a deictic expression like *there* and a deictic gesture (like pointing, a head nod or directed gaze; Levinson, 2006), which, together, serve to identify a location in a given situation. This condition, however, may be and, as will be argued in this paper, is, in fact, incomplete. While a complex form might be a verbal plus a kinetic element, it might also be a verbal element plus a prosodic pattern. To show that the second combination is also a possible manifestation of a multimodal construction, it needs to be shown that morphosyntax and prosody are two different modes, each contributing independently to the meaning of the construction. Alternatively, it might be assumed that prosody is yet another aspect of unimodal constructions, on a par with their phonological properties. The review provided in Section 3 will rule out this alternative viewpoint.

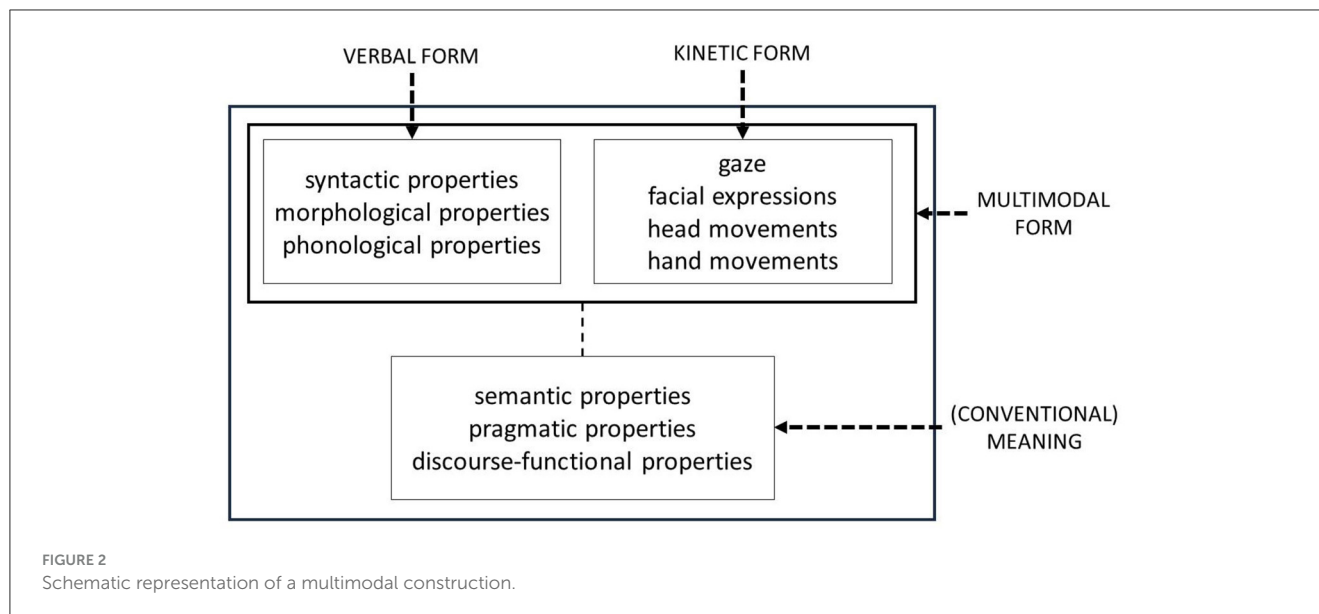
The second condition Ziem (2017) puts forward runs as follows:

- (b) Multimodal constructions manifest themselves either as inherently multimodal units or as entrenched cooccurrences of a verbal and a kinetic element (as opposed to constructions solely realized in a multimodal way).

This condition indicates that there are two kinds of multimodal constructions, which need to be kept distinct from incidental cooccurrences of e.g., a construction and a gesture (see also Hoffmann, 2017). The first kind of multimodal construction is inherently multimodal, i.e., it is non-predictable in some way. This holds for the combination of a deictic expression and a deictic gesture: The deictic expression remains incomplete in meaning (at least in some of the cases) unless it is used with deictic gesture. The second kind of multimodal construction follows from the usage-based premise that an expression can be fully predictable and still be a construction when it occurs with sufficient frequency. Schoonjans (2018), for example, could show that the German particle *einfach* cooccurs with a head shake in 24% in his corpus. Zima (2017) could show that [all the way from X PREP Y] is produced with a gesture in 80% of cases. And Uhrig (2022) could show that verbs of throwing are, on average, accompanied by a gesture in 54% of cases (with 66% for *fling* but only 42% for *lob*). Even though these corpus studies attest statistically significant cooccurrences of morphosyntactic and kinetic elements, they could only provide indirect evidence that this statistical significance can be equated with practical significance, i.e., show that these multimodal realizations constitute cognitive units. Therefore, in Section 4, the present paper will provide some evidence that language users actively make use of the prosodic mode to disambiguate (multimodal) constructions by reporting on a forced-choice experiment using the construction *Tell me about it*.

The present paper is not the first trying to bring together Construction Grammar and prosody. The past decade has also seen a rise in studies researching the prosody-syntax interface from a Construction Grammar perspective, but did so independently, i.e.,

<sup>1</sup> The other two conditions follow from the first two and therefore do not need explicit attention. The third condition specifies what should not be considered a multimodal construction (e.g., a construction only realized multimodally) and the fourth condition states that multimodal constructions need to be part of the constructional network of a language, i.e., a network that covers the relevant knowledge a speaker of that language needs for understanding.



without referring to multimodal constructions. In the Introduction to their edited volume on Prosody and Construction Grammar, Imo and Lanwer (2020) summarize possible synergies. One possibility is the existence of prosodic constructions, i.e., assemblies of prosodic features that convey a particular meaning (relatively) independent of the words that are used with it. These prosodic constructions combine with morphosyntactic constructions in an *ad hoc* manner if their functions are compatible. Prosodic constructions have been proposed for French (Marandin, 2006), Persian (Sadat-Tehrani, 2010), Spanish (Gras and Elvira-García, 2021), and English (Ward, 2019). Another possibility is that prosodic properties, if recurring, can be part of the formal side of the (unimodal) construction. This was proposed for the reactive *what-x* construction (*What mince pies?*), which reacts to something in the preceding turn by another speaker and needs to be prosodically integrated (Pöldvere and Paradis, 2020). And, finally, a third possibility is that prosody and morphosyntax interact in a meaningful way such that a construction would be incomplete without considering both components and none of the two components constitute independent constructions. This seems to be the case for German appositive structures (e.g., *der Spitzenkoch Tim Mälzer*, English *the top chef Tim Mälzer*), as evidenced in Lanwer (2020). Even though this is not made explicit, this possible relation between prosody and Construction Grammar fits the definition of a multimodal construction with the only exception that “kinetic” form needs to be replaced by “prosodic” form. Figures 3–5 summarize all possible configurations.

In a nutshell, the present paper aims to show that there are multimodal constructions that consist of a syntactic and a prosodic form, which combine to convey one meaning. To do so, evidence for a prosodic mode (in English) will be reviewed to show that, in principle, prosody and morphosyntax (or rather the phonological properties of morphosyntactic elements) are two different modes. Moreover, a forced-choice experiment will be reported on, which shows that certain prosodic forms are not just used incidentally, but that they are part of language users’ knowledge.

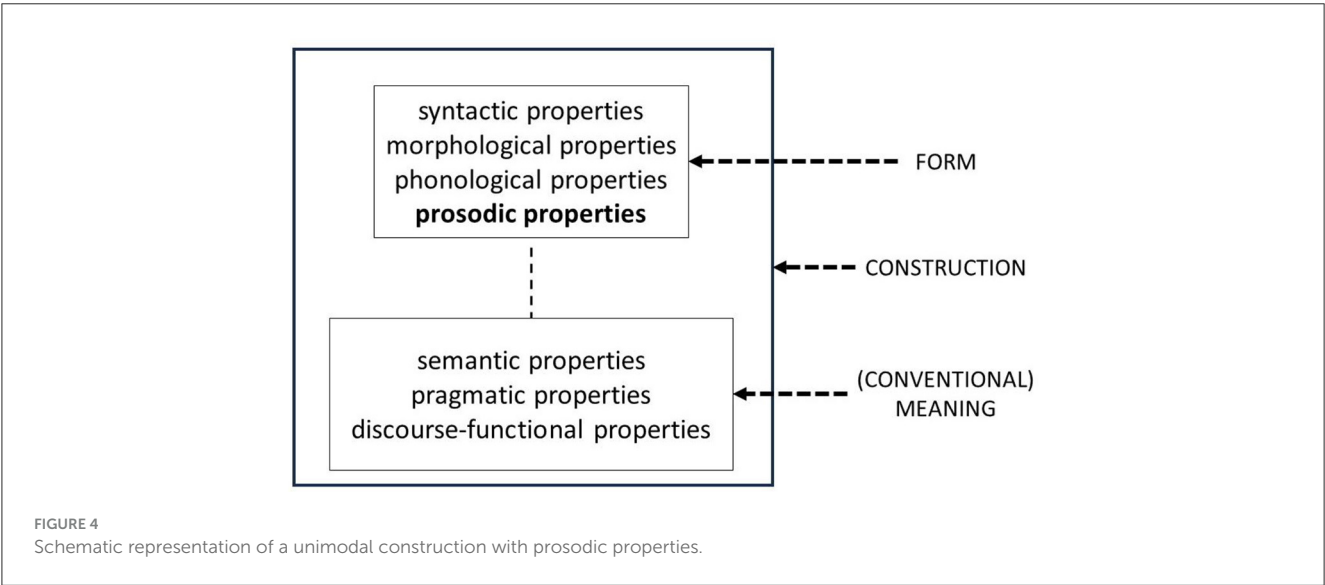
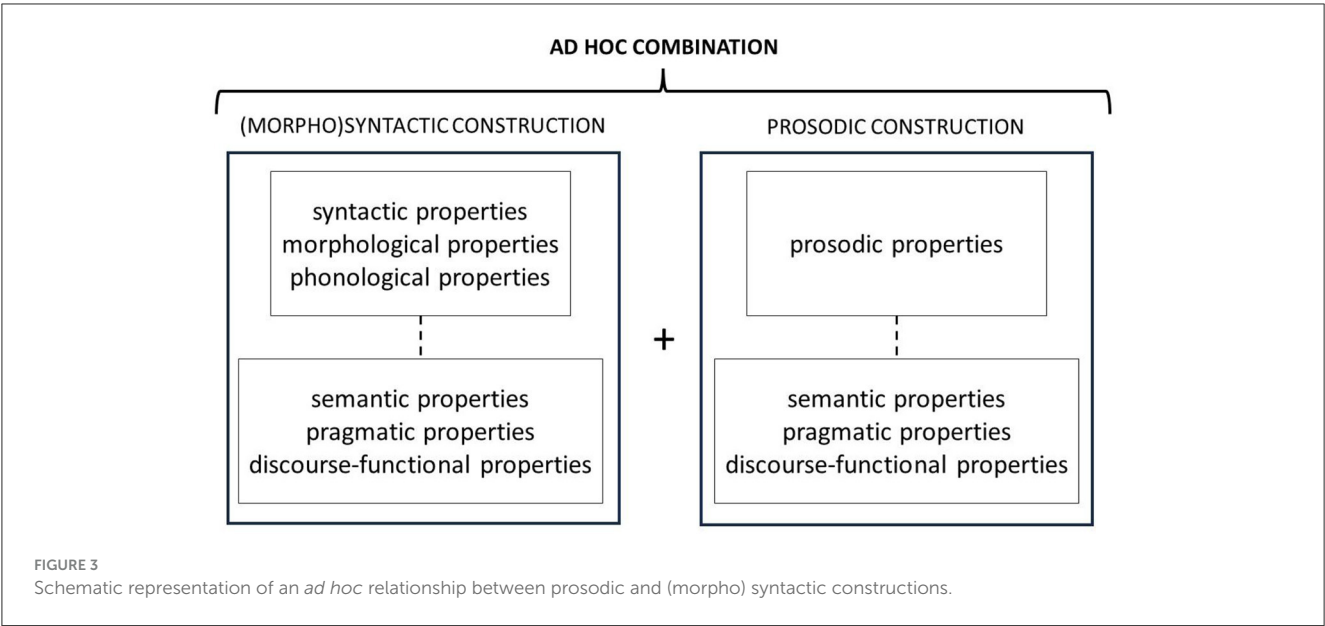
### 3 Evidence for a prosodic mode in English

There are many definitions of the term mode and some of them equate mode with sensory channel. Such a, often pre-theoretical, notion of mode might be one of the reasons why prosody has been largely neglected in usage-based, multimodal approaches to Construction Grammar. From such a view, prosody and spoken language belong to the same mode and, thus, need not be part of multimodal analyses. The present paper, however, will use the notion of semiotic mode, which is prevalent in multimodality research. More specifically, the paper will make use of the definition of semiotic mode as proposed by Bateman and colleagues (Bateman, 2011, 2022; Bateman and Wildfeuer, 2014; Bateman et al., 2017).

Bateman defines a semiotic mode as “a three-way layered configuration of semiotic distinctions developed by a community of users in order to achieve some range of communicative or expressive tasks” (Bateman, 2022, p. 68). The first layer of the semiotic mode is the material substrate, i.e., “the ‘stuff’ which is used when making meaning” (Bateman and Wildfeuer, 2014, p. 181). In other words, semiotic agents manipulate the material to communicate. The second layer is the form side of the mode. The form consists of categories derived from the (noisy) material that are, by convention, used to distinguish meanings. These forms can be simple or complex. And, finally, the third layer of the semiotic mode is that of discourse semantics, i.e., the meaning contribution of the mode in relation to its surroundings. The following subsections will show if and to what extent (spoken) morphosyntax and prosody differ along these lines.

#### 3.1 The material substrate

From an articulatory perspective, the material substrate of spoken English morphosyntax is part of introductory knowledge



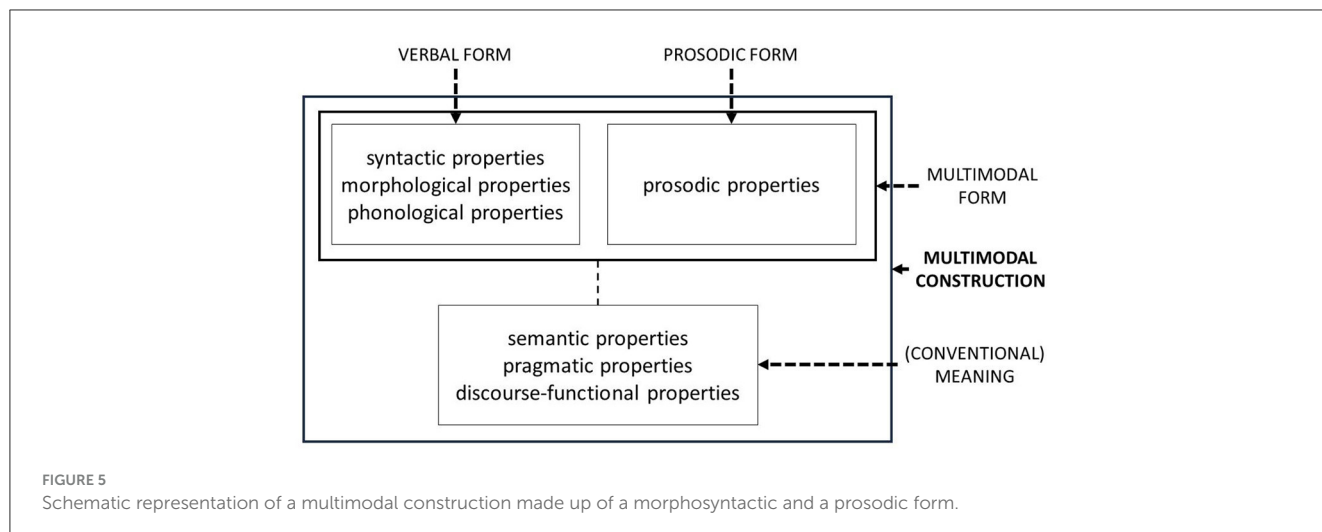
in linguistics. Speakers use the air stream coming from the lungs and manipulate this air stream with the help of different, active and passive, articulators to create sounds. One main active articulator is the vocal folds, which can produce voiced sounds when vibrating and voiceless sounds when not vibrating. The other articulators of English sounds are mainly found in the oral cavity: the lips, the teeth, the tongue, the alveolar ridge, the hard and the soft palate (also called velum) as well as the uvula (depending on the variety of English spoken). Acoustically, this manipulation of the airstream results in different shapes of the sound waves produced. For example, plosive sounds are characterized by a silent period and a sudden release burst, fricatives by a strong turbulence noise and vowels by energy peaks at certain frequencies (also known as first and second formants), to name but a few.

The articulatory mechanisms behind prosodic features in English (to be discussed below) partially overlap with that of the sounds of English. The most central prosodic features—pitch,

loudness, and duration—are manipulated largely with the help of the diaphragm and the vocal folds. The diaphragm is a large muscle below the lungs that controls breathing and thus, the airstream. The greater the airflow, the louder the speech tends to get. The diaphragm is also involved in producing (English) speech sounds, because, when there is no airflow, no sounds can be produced.<sup>2</sup> Technically, speakers may also “speak from their throats,” i.e., without support from the diaphragm, but even that has respiratory constraints. Still, even though the diaphragm is involved in the production of speech sounds, it does not have an influence on the perception of these sounds as phonemes. *A/l*/is *a/l*/, no matter whether it is loud or quiet. Acoustically, with greater airflow, the pressure the sound signal exerts on the surrounding particles is

<sup>2</sup> Languages other than English have non-pulmonic sounds, i.e., sounds where the airflow does not come from the lungs, but these will not be considered here.





higher. The other main articulator in prosody is the vocal folds, which are responsible for pitch production. The speed with which they vibrate correlates with the fundamental frequency ( $f_0$ ) of the sound produced. The faster they vibrate, the higher the sound is perceived. As outlined above, the vocal folds are also involved in sound production. However, even though the articulator is the same, it does two different things here. For sound production, it is important to either let the vocal folds vibrate or not. For pitch, what matters is the speed with which they vibrate. From an acoustic perspective, higher frequency of vibration causes the sound waves to oscillate faster, too.

All in all, what can be seen from this necessarily brief overview is that sounds (as the building blocks of spoken morphosyntax) and prosodic features are produced by different parts of the articulatory system. This means that they can be (and are) manipulated independently in the meaning-making process and, thus, also can take on different forms.

### 3.2 Form

Regarding the form of spoken English morphosyntax, the paper will only consider phonological categories, since these are most central for the present argument. The phonological features that serve meaning-distinguishing purposes in English are the state of the glottis, the manner of articulation, and the place of articulation for consonants, and the positioning of the tongue and duration for vowels. For English vowels, further meaning-distinguishing features have been proposed, either in addition or substituting duration, namely muscular tension and position of the lips. In any case, features like these enable language users to distinguish categories such as /b/ and /p/ (state of the glottis), /b/ and /m/ (manner of articulation), /b/ and /d/ (place of articulation), /i:/ and /u:/ (position of the tongue) as well as /i:/ and /ɪ/ (duration, but also position of the tongue).

For prosody, features that serve meaning-distinguishing purposes include, at least, the “big three” pitch (the perceptual correlate to fundamental frequency), loudness (the perceptual

correlate of the pressure of the sound signal), and aspects of timing (such as speaking rate, articulation rate or pauses). These features enable the language user to perceive categories such as rising and falling intonation (pitch), loud and quiet speech (loudness) as well as fast and slow speaking tempo (timing). These three often work together to form prosodic constructions, i.e., configurations of prosodic forms that convey a particular meaning independent of the words used (see Section 3.3. below for examples). There are further prosodic features, such as voice quality (nasality, creakiness) and articulatory precision, but these seldom serve meaning-distinguishing functions on their own. In sum, there is some overlap regarding the meaning-distinguishing features of spoken morphosyntax and prosody, since (vowel) duration and timing are both time-related features, but other than that, the features can clearly be distinguished from one another. What is more, even though vowel duration and timing seem to correlate, language users are able to distinguish the two nonetheless. Just consider a word like *bit*. Its vowel, /ɪ/, is short in duration, but the meaning of the word does not change if it is pronounced in a slow manner (which is the case, of course, because no two words in English are ever distinguished by vowel duration alone) as long as the contrast with other vowels of a similar quality is maintained.

An interesting exception might be stress placement. There are words in English that only differ by word stress, e.g., *differ* /ˈdɪfə/ or /ˈdɪfər/ and *defer* /dɪˈfɜ:/ or /dɪˈfər/. The acoustic correlates of stress in English include, among others, pitch, loudness and timing (see e.g., Fry, 1955, 1958; Lieberman, 1960), i.e., the “big three” mentioned above. Examples like *differ* and *defer* blur the lines between meaning-distinguishing features that are relevant for morphosyntax and those for prosody. Therefore, one could treat them as counterevidence that prosody is an independent mode because a prosodic configuration that language users perceive as word stress serves morphosyntactically relevant functions. Likewise, it could be argued that words like *differ* and *defer* are, in fact, multimodal constructions combining a phonological (e.g., /dɪfə/) and a prosodic form (e.g., /ˈσσ/) for *differ*. It is outside the scope of the present paper to provide evidence for one or the other claim. Still, the argument put forward in the following clearly favors the second option.

### 3.3 Discourse meaning

From a Construction Grammar perspective, all morphosyntactic units of interest, i.e., constructions, carry meaning per definition (although this is not uncontroversial, see e.g., Fillmore et al., 2012 on constructions without meaning). Therefore, there is no need to discuss the meaning of these.

The more interesting question is rather whether prosodic forms, independent of the words that are used with them, carry meaning. There is, in fact, quite some evidence for the existence of prosodic constructions. Prosodic constructions have been identified for Spanish (Elvira-García, 2019; Gras and Elvira-García, 2021), German (Neitsch and Niebuhr, 2019; Niebuhr, 2019), French (Marandin, 2006), Persian (Sadat-Tehrani, 2010), and most notably for the present purposes, English (Ward, 2019). One of the prosodic constructions attested for English, the *consider this* construction, will be reviewed in more detail, because it is one of the constructions that is understood best. This prosodic construction was first described in Liberman and Sag (1974) and is attested both experimentally (Kurumada et al., 2012) and with the help of corpora (Hedberg et al., 2003; Ward, 2019). Its formal features are illustrated in Figure 6. While most of its formal descriptions focused on the pitch movements only, recent advances show that it consists formally of three parts: The first is a region that is high-pitched, loud and slow, to be seen on the word LOOKS in Figure 6. The second is a region of level pitch, which can be seen on *like a ze-* in Figure 6. And, third, another high-pitched region, visible on the last syllable *-bra* in Figure 6 (Ward, 2019, p. 5–24). Functionally, it marks some kind of contradiction or contrast, a piece of information that is offered to the hearer for further consideration. Thus, the syntactic string *It looks like a zebra* uttered with the prosodic pattern described above implies that even though the animal in question might resemble a zebra, it is actually some other animal (Kurumada et al., 2012). There is compelling evidence that this form-function pairing is indeed conventionalized in American English: Corpus studies suggest that this prosodic form is more often than not used with contradictions (Hedberg et al., 2003; Ward, 2019) and experimental evidence suggests that language users favor a “no zebra” interpretation when presented with an utterance like depicted in Figure 6 (Kurumada et al., 2012). What is more, Liberman and Sag (1974) even argue that “without having any idea of the content of his utterance, we know from the melody performed ... that [the speaker] objects in some way” (422), i.e., that the prosodic form has an independent meaning. This independent contribution to the discourse semantics of prosody is probably the most convincing piece of evidence that prosody is an independent semiotic mode.

### 4 Entrenching prosodic information: *Tell me about it*

Section 3 argued that prosody is best seen as an independent semiotic mode. For the discussion on the relation between prosody and morphosyntactic constructions this means that prosodic properties cannot be analyzed on a par with other properties of morphosyntactic construction but need independent consideration. Section 3.3, in particular, has shown that there

are prosodic constructions, like the *consider this* construction, that may combine with morphosyntactic constructions in an *ad hoc* manner to form a multimodal construct. In what follows, the paper will present some evidence for a genuinely multimodal construction, i.e., a construction with both entrenched prosodic and morphosyntactic properties. The construction under consideration is called stance-related *Tell me about it* and will be contrasted with another, formally similar construction, i.e., requesting *Tell me about it*.

### 4.1 Requesting and stance-related *Tell me about it*

Formally, requesting and stance-related *Tell me about it* (henceforth TMAI) are morphosyntactically similar. While formal variations for the stance-related construction can be found (e.g., *Tell me more* or *Tell me more about it*), these are rare and *Tell me about it* seems to be the preferred variant as this is the only form that is listed in dictionaries (e.g., in the Oxford English Dictionary Online, Tell, 2023). Functionally, the two TMAI constructions fulfill different, non-overlapping functions. Requesting TMAI is used to request information as is illustrated in Example (2).<sup>3</sup>

- (2) “sci-fi thriller” (simplified)  
 A: I know she also has a sci-fi thriller. Arrival.  
 B: Uh-huh.  
 A: Tell me about it. Is it worth seeing?  
 B: Absolutely.  
 (2016-09-25\_0832\_US\_KNBC\_Access\_Hollywood, 29:41–29:48).

In Example (2), speaker A introduces a referent, i.e., a science fiction thriller called *Arrival*. After speaker B’s brief backchannel, speaker A encourages speaker B to provide more information on this film using TMAI and specifies the preferred continuation to be an evaluation (i.e., *Is it worth seeing?*). Speaker B then provides the requested information. As can be seen in this example, requesting TMAI usually initiates speaker transition. This transition need not occur directly after issuing TMAI, but constitutes what Sacks et al. (1974) call a transition-relevance place. Moreover, the next turn is expected to be an informing sequence, providing some more information on the referent that was introduced shortly before.

Stance-related TMAI fulfills completely different functions as is illustrated in Example (3).

- (3) “we’re all getting older” (simplified)  
 A: We’re getting older. We’re all getting older. So ...  
 B: ((laughs)) T- Tell me about it.  
 A: ((laughs))  
 (2021-11-26\_0600\_US\_KNBC\_Dateline\_NBC, 03:39–03:44).

<sup>3</sup> All examples of TMAI come from the NewsScape Library of International Television News, an archive of televised discourse (Steen and Turner, 2013). At the end of each example, the name of the source file and the relevant times are provided. Video snippets of the examples are provided on OSF: [https://osf.io/2sq7h/?view\\_only=746f3703bbde4236b832b34234d51beb](https://osf.io/2sq7h/?view_only=746f3703bbde4236b832b34234d51beb).

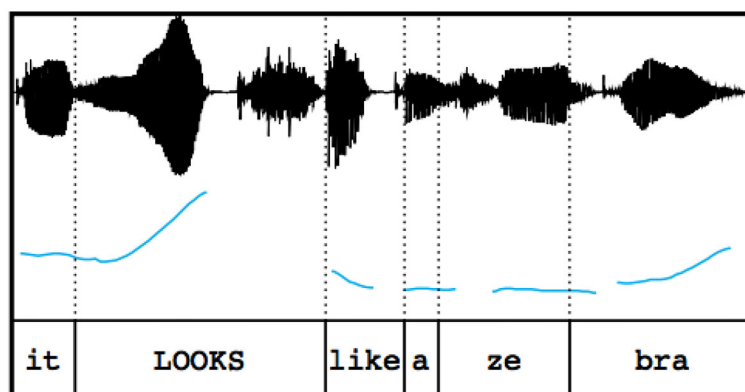


FIGURE 6  
Waveform and pitch contour of the “consider this” construction (taken from Kurumada et al., 2012).

In Example (3), speaker A makes an observation (*we’re getting older*), which many people find saddening. This seems also to hold for speaker A since he repeats this utterance, slightly modifying it (*we’re all getting older*). Speaker B reacts to this observation, at first, with laughter and then with stance-related TMAI. This construction expresses an affective stance, i.e., a saddening view on aging. Likewise, it expresses epistemic authority. Speaker B is, apparently, older than speaker A and thus claims to be more knowledgeable person on this matter. Crucially, stance-related TMAI neither necessitates speaker transition nor an informing sequence. Speaker A reacts with laughter to speaker B uttering TMAI and the conversation is cut at this point.

It could be argued that both TMAI constructions are ambiguous and are only disambiguated in predictive context. However, in a corpus study using the multimodal *NewsScape Library of International Television News* (Steen and Turner, 2013), Lehmann (2023) showed that stance-related TMAI, when compared to requesting TMAI, is produced, more often than not, with raised eyebrows, averted gaze, smiling, some kind of head movement (often nods, shakes or tilts) and a slower speaking rate. This is illustrated with frame grabs of Example (3), which are provided in Table 1.

As can be seen Table 1, before uttering stance-related TMAI, speaker B looks at his interlocutor, already smiling. At the onset of TMAI, he turns his head (line 2) to the left and avoids eye contact with the recipient. In addition, he raises his eyebrows and continues smiling (see also line 3). Only after finishing uttering TMAI, on the last syllable, he turns his head orientation and his gaze back toward his interview partner. The duration of TMAI in Example (3) is 667 ms, which corresponds to a speaking rate of 7.4 syllables per second. This is very close to the mean speaking rate of stance-related TMAI in face-to-face interactions, which is 7.48 syllables per second, whereas requesting TMAI is faster in these contexts, with a speaking rate of 8.44 syllables per second (see Lehmann, in press).

All of these visual as well as prosodic properties of stance-related TMAI were shown to be statistically significant (Lehmann, 2023), but as was argued above, some Construction Grammarians claim that statistical significance need not be equated with practical significance. Therefore, both visual and prosodic properties of

TMAI were put to the test in a forced choice experiment to provide evidence that language users indeed draw on these properties when interpreting an instance of TMAI.

## 4.2 Putting the multimodal properties of *Tell me about it* to the test

### 4.2.1 Method





#### 4.2.1.1 Participants

The participants in this experiment were 25 adult native speakers of American English, who were recruited via Prolific Academic (Palan and Schitter, 2018). They were rewarded £4.50 for their participation. In addition, 18 adult advanced learners of English participated. These were students of the study program *English-speaking Cultures* at the University of Bremen, Germany. To be admitted to this study program, students need to have a command of English at level B2 (“independent user”) of the Common European Framework of Reference for Languages (Council for Cultural Co-operation, Education Committee, and Modern Languages Division, 2001), but many of them self-reported to know English on a C1 level (“proficient user”). They participated for course credit.

#### 4.2.1.2 Procedure

The participants were requested to complete an online forced choice experiment, which had been designed with SoSci Survey (Leiner, 2021). In the instructions to this experiment, the participants were introduced to the two uses of TMAI, named *requesting information* and *ironic rejoinder*. This was done to make sure that the non-native speaker understand the task (in case they did not know TMAI could also be used in a stance-related way) and to introduce the two response options in the experiment. The label *ironic rejoinder* was preferred over the label *stance-related* in the experiment because the *Oxford English Dictionary* defines stance-related TMAI this way (Tell, 2023). The participants were told that they would see and/or hear a speaker uttering TMAI and that their task was to guess whether this utterance is requesting information or an ironic rejoinder.

TABLE 1 Frame grabs of an extract of example (3).

Line	Speaker	Utterance	Frame grab
1	A	So...	
2	B	t-	
3		Tell me about	
4		It	

4.2.1.3 Stimuli

The experiment consisted of 69 stimuli in total. All of these were selected observations of the corpus study from [Lehmann \(2023\)](#). These observations were presented in four different conditions.

In the first condition, called “context condition,” the participants were presented with TMAI with what was considered sufficient sequential context to disambiguate TMAI with the help of this context. This served as the reference condition. In the second



TABLE 2 Overview on the stimuli used in the experiment.

Condition	Description	Anticipated interpretation
Context	TMAI embedded in sequential context	Requesting ( $N = 5$ )
		Stance-expressing ( $N = 5$ )
Multimodal	Stand-alone TMAI Visual and acoustic information provided	Requesting ( $N = 5$ )
		Stance-expressing ( $N = 4$ )
		Ambiguous ( $N = 9$ )
Visual	Stand-alone TMAI No acoustic information Pace slowed down	Requesting ( $N = 5$ )
		Stance-expressing ( $N = 5$ )
		Ambiguous ( $N = 11$ )
Acoustic	Stand-alone TMAI No visual information	Requesting ( $N = 5$ )
		Stance-expressing ( $N = 4$ )
		ambiguous ( $N = 11$ )

condition, called “multimodal condition,” the participants could both hear and see a speaker uttering TMAI, but without further sequential context. In the third condition, called “visual condition,” the participants saw a speaker uttering TMAI, but they could not hear this person. Since these video snippets were extremely short with less than a second and some online video players have a time lag, the videos were played in slow motion. The participants were informed about this. Furthermore, to facilitate speaker identification in case there was more than one speaker visible, the videos were edited to such an extent that only the speaker of TMAI was visible. Finally, in the fourth condition, called “acoustic condition,” the participants were provided with an audio recording of a speaker uttering TMAI only. Within, but not between these conditions, stimuli rotated.

The stimuli were further selected regarding their anticipated interpretation. The statistical model that was fitted for the corpus data in Lehmann (2023) makes clear predictions about how participants should interpret these stimuli, if the results were of practical significance. Thus, stimuli were selected according to the visual and/or prosodic features that the speakers used during the utterance. That is, some stimuli were selected as either prototypically requesting or stance-related uses of TMAI, when they displayed the properties that the statistical model predicted. Vice versa, some of the stimuli were selected as ambiguous stimuli when they displayed conflicting properties, e.g., when the speaker raised their eyebrows (a property of stance-related TMAI) but continued looking at the recipient (a property of requesting TMAI).

Table 2 gives an overview on the stimuli used in the experiment.

## 4.2.2 Statistical analysis

The results of the forced choice experiment were analyzed with R (R Core Team, 2022). With the help of the *glmer* function of the *lme4* package (Bates et al., 2015), a generalized linear mixed-effects

model was fitted. The correctness of the response (i.e., whether the response was in line with the actual construction) was treated as the dependent variable. Initially, participant, language proficiency, stimulus, and construction were entered as random intercepts, while condition and anticipated interpretation were entered as fixed effects. This led to problems with convergence due to its complexity. An inspection of the initial model with the *summ* function of the *jtools* package (Long, 2022) showed that language proficiency and participant were negligible effects and were, thus, removed from the model. No problems with convergence occurred thereafter. The *summ* function was used to summarize the fitted model, including the computation of confidence intervals, and the *ggplot2* package (Wickham et al., 2023) as well as the *sjPlot* package (Lüdtke, 2023) were used to visualize the fitted model.

## 4.2.3 Results

Figure 7 shows the overall distribution and central tendencies of correct responses for the different stimuli across conditions.

Figure 7 suggests that, overall, the participants were successful at guessing the meaning of TMAI based on visual and/or acoustic cues alone, given that the median ratio of correct guesses for the unambiguous stimuli is higher than 0.75. Figure 7 also suggests that, when compared to the context condition, participants seemed to have difficulties with the ambiguous stimuli, but neither the requesting nor the stance-related ones, except for five stimuli which score lower than 0.75, three of which in the visual condition and two in the acoustic condition.<sup>4</sup> In general, participants perform worse in the visual and the acoustic condition than in the multimodal condition. In these two conditions, the ambiguous stimuli seem to pose the greatest difficulties to the participants, as expected.

Table 3 provides a summary of the fitted model and Figure 8 shows the odds ratios of the model terms (condition and anticipated interpretation).

With a pseudo- $R^2$  of 0.64 for the total effects and a pseudo- $R^2$  of 0.36 for the fixed effects, the model summarized in Table 3 explains a good amount of variance in the responses obtained. It shows that the participants were significantly worse at guessing the meaning of TMAI in the multimodal (with  $p = 0.04$ , OR = 0.13), visual (with  $p < 0.001$ , OR = 0.02) and acoustic condition (with  $p < 0.001$ , OR = 0.002) when compared to the context condition. It further shows that there is no significant difference between guessing requesting and stance-related TMAI correctly (with  $p = 0.33$ , OR = 2.09), but the ambiguous stimuli contribute to the model with borderline significance (with  $p = 0.06$ , OR = 0.35), suggesting that most incorrect guesses were due to the ambiguous stimuli, but not entirely.

4 There seem to be at least two reasons why the participants scored low in correctness for these prototypical stimuli. One reason might be the timing of TMAI and the visuals. That is, for some visual stimuli, some important visual displays (gaze aversion, raised eyebrows, and smiling) occurred right before, but not during the speaker uttered TMAI. This non-synchrony might have affected the speakers' choices. Another reason might be that the model reported in Lehmann (2023) is incomplete. It seems that, while the duration of TMAI is a good predictor, it is not the only one.



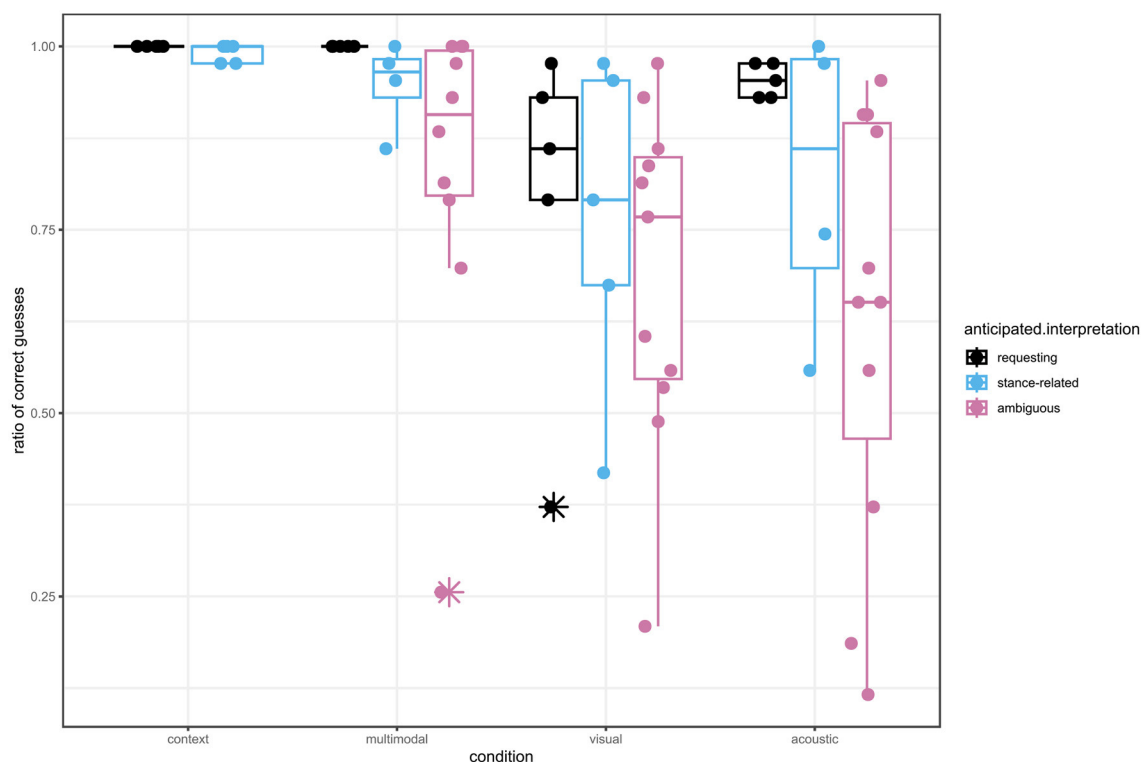


FIGURE 7

Grouped boxplots with jitter of correct responses regarding the anticipated interpretation across conditions. The asterisk indicates outliers.

#### 4.2.4 Discussion of the forced-choice experiment

The experiment reported above shows that prosody alone can disambiguate TMAI if the prosodic features that are associated with the construction are displayed, i.e., the speaking rate in this case. If TMAI is ambiguous regarding its speaking rate and hearers lack other pieces of information, they seem to have difficulties in guessing its meaning. Vice versa, if the speaker produces TMAI with a slower speaking rate, hearers are more likely to understand this as stance-related TMAI, even if there are no further features available. Interestingly, the results also suggest that hearers use prosodic information alone to disambiguate TMAI about as accurate as they use visual information alone. This observation might suggest that the strength of association between prosodic properties and the construction is comparable to the one between visual properties and the construction.

Technically, these observations can be explained in two ways. One explanation is that slow speaking rate is an independent prosodic construction. Niebuhr (2010), for example, has shown that lengthened consonants correlate with negative sentiment in German. The same could be true for English stance-related TMAI. Informal observations of TMAI, however, suggest that it is not the lengthening of the consonants alone that result in a slower speaking rate, but also the lengthening of the vowels. At the same time, speaking rate alone does not explain all the findings observed in the experiment. There are quite a few stimuli that were neither slow nor fast (i.e., ambiguous), which posed no difficulties to the participants. This suggests that there might be more, albeit undetected, prosodic features associated with TMAI. Given that,

it is possible that there is a (complex) prosodic construction that is often used with stance-related TMAI, but, at the moment, there is only scarce evidence for that. The other way to explain the findings of the experiment is to assume that the slow speaking rate is part of the stance-related construction, forming a multimodal construction. If there is, indeed, no prosodic construction that can be identified, and given that prosody is a mode, then stance-related TMAI must be considered a multimodal construction with morphosyntactic and prosodic (and, possibly, visual) features. Even if future studies show that there is a prosodic construction such as “slow speaking rate,” both the frequency with which it is used with stance-related TMAI and the apparent use of this construction to disambiguate TMAI would speak in favor of treating TMAI as a multimodal construction from a usage-based perspective.

## 5 Conclusions

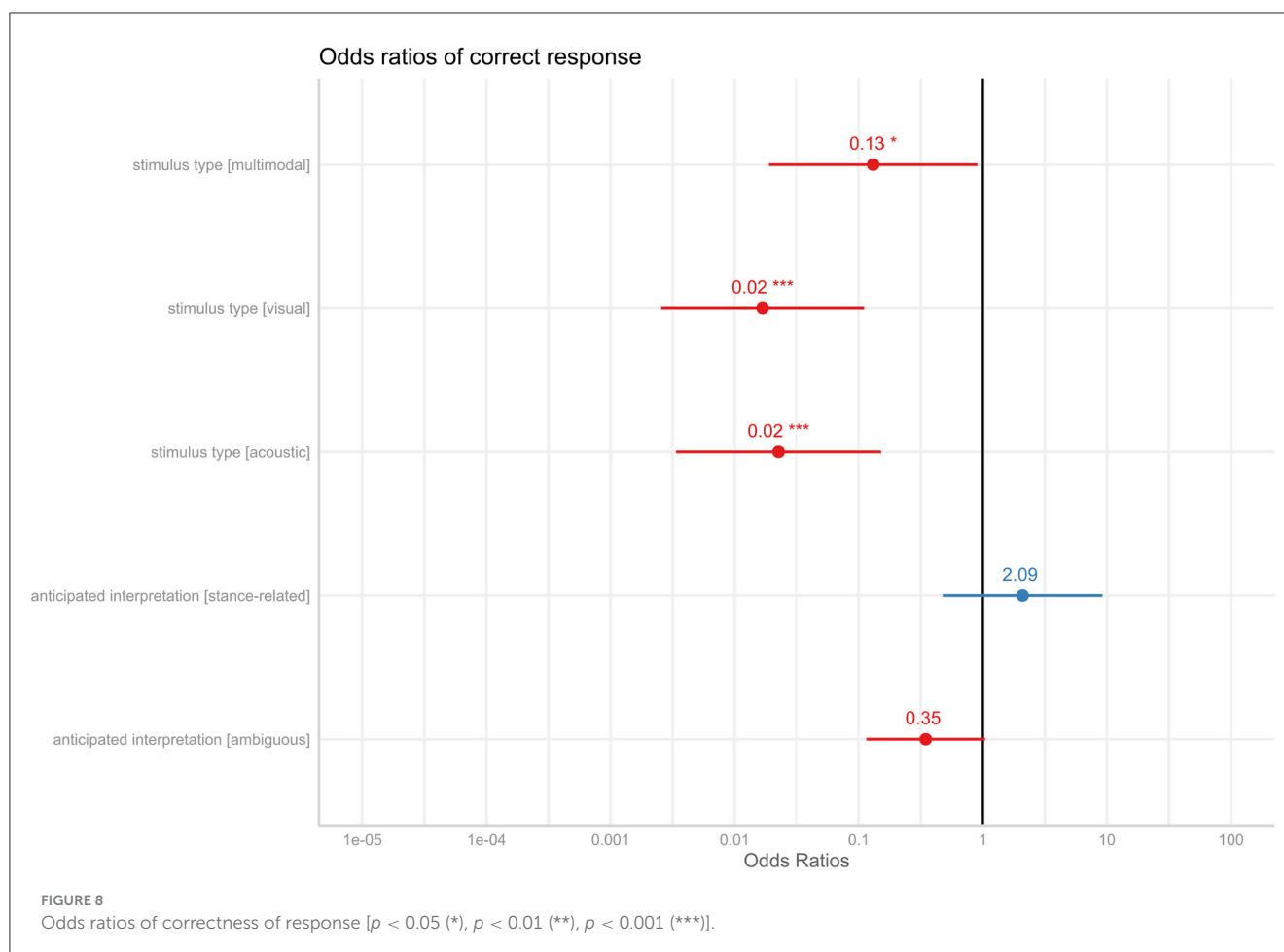
The present paper had two objectives. The first objective was to show that prosody and morphosyntax are two independent semiotic modes with distinguishable differences in material and form as well as independent contributions to the discourse semantics. It could be shown that the aspects of the sound stream that are relevant for spoken morphosyntax are not the same as the aspects that are relevant for prosody. Using these different aspects, hearers transform the input from the sound stream to either arrive at categories like /p/, /m/ or /e/ (spoken language) or high pitch, loud speech, and/or fast

TABLE 3 Summary of the fitted model for correct responses.

Model info:					
Observations: 3010					
Dependent Variable: correctness					
Type: Mixed effects generalized linear regression					
Error Distribution: binomial					
Link function: logit					
Model fit:					
AIC = 1,997.97, BIC = 2,046.05					
Pseudo-R <sup>2</sup> (fixed effects) = 0.36					
Pseudo-R <sup>2</sup> (total) = 0.64					
Fixed effects:					
	Est.	2.5%	97.5%	z val.	P
(Intercept)	5.87	3.46	8.27	4.79	<0.001
Multimodal	−2.03	−3.97	−0.10	−2.06	0.04
Visual	−4.09	−5.97	−2.20	−4.25	<0.001
Acoustic	−3.79	−5.69	−1.89	−3.91	<0.001
Stance-related	0.74	−0.74	2.22	0.98	0.33
Ambiguous	−1.06	−2.16	0.04	−1.89	0.06
Random effects:					
Group	Parameter	Std. dev.			
Stimulus	(Intercept)	1.25			
Construction	(Intercept)	0.97			
Grouping variables:					
Group	# groups	ICC			
Stimulus	70	0.27			
Construction	2	0.16			

speech (prosody). These categories are then combined to form meaningful structures like *It looks like a zebra* (spoken language) or (contextually meaningful) assemblies conveying “consider this” (prosody), and they do so largely independent of one another. Since spoken language and prosody differ in all three layers of the semiotic mode, they must be considered independent. For constructional analyses, this means that prosody cannot be represented on a par with other, morphosyntactic and phonological, properties. Rather, it needs its own place. This place could take on the form of a prosodic construction (in case the prosodic configuration has an independent meaning) or of being part of a multimodal construction (in case the prosodic configuration has no independent meaning). Such a view on prosody strengthens the multidimensional network approach to language-related knowledge, which assumes that constructions are interrelated by various kinds of associations (Diesel, 2023). Prosodic constructions as well as multimodal constructions are prime examples of such a network of (cross-modal and multimodal) associations.

The second objective of the present paper was to provide evidence for a multimodal construction consisting of, at least, a morphosyntactic and a prosodic form. Both corpus and experimental evidence suggest that the stance-related use of *Tell me about it* is a likely candidate for such a multimodal construction. Regarding its prosodic form, stance-related *Tell me about it* is slower in tempo than its requesting counterpart. When language users are provided with nothing but this difference in tempo (i.e., they lack other clues like sequential context or visuals), they use this prosodic feature to disambiguate *Tell me about it*. In other words, this knowledge on the two uses of *Tell me about it* must be stored in the language users’ minds in some way. Stance-related *Tell me about it* thus fulfills Ziem’s second condition of multimodal constructions, because it cannot be considered a construction that is “solely realized in a multimodal way,” but the paper has shown that it is an entrenched cooccurrence of a verbal and a prosodic form. In conclusion, the evidence presented in this study on *Tell me about it* is strongly suggestive of the existence of multimodal constructions. As a consequence, the role



prosody plays in forming them needs more systematic attention in constructional analyses.

From a methodological perspective, the present paper could show that a triangulation of corpus and experimental evidence is valuable because it was able to shed light on both the production and the comprehension side of language and, in doing so, draw a complementary picture of prosody and multimodal constructions. However, the present study suffers from obvious limitations that require further systematic attention in future studies. One limitation is the low number of participants in the forced-choice experiment and the missing demographic information. From a usage-based perspective, the constructional network (including multimodal and prosodic constructions) is dynamic and, therefore, can vary for certain demographic groups. This aspect is not reflected in the present study and needs to be addressed in the future. In addition, future research also needs to address the role prosody plays in the constructional network in more detail. Studies that explore prosodic and multimodal constructions could identify the exact (inter)relations and associations between different types of constructions and, thereby, provide an answer to the question if multimodality is a central or a peripheral aspect of grammar.

## Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and

accession number(s) can be found at: [https://osf.io/2sq7h/?view\\_only=746f3703bbde4236b832b34234d51beb](https://osf.io/2sq7h/?view_only=746f3703bbde4236b832b34234d51beb).

## Ethics statement

Ethical approval was not required for the study involving human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants in accordance with the national legislation and the institutional requirements.

## Author contributions

CL: Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Projektnummer 491466077.

## Acknowledgments

My thanks go to John Bateman. John was the best mentor you can imagine during my time at the University of Bremen. Without him, this paper would be less rigorous and less advanced in almost every respect.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

- Aarts, B., Bowie, J., and Popova, G. (2019). "Introduction," in *The Oxford Handbook of English Grammar*, eds. B. Aarts, J. Bowie, and G. Popova (Oxford: Oxford University Press), 1.
- Bateman, J. (2011). "The decomposability of semiotic modes," in *Multimodal Studies: Exploring Issues and Domains*, eds. K. O'Halloran and B. Smith (New York, NY: Routledge), 17–38.
- Bateman, J. (2022). Growing theory for practice: empirical multimodality beyond the case study. *Multimodal Commun.* 11, 63–74. doi: 10.1515/mc-2021-0006
- Bateman, J., and Wildfeuer, J. (2014). A multimodal discourse theory of visual narrative. *J. Pragmat.* 74, 180–208. doi: 10.1016/j.pragma.2014.10.001
- Bateman, J., Wildfeuer, J., and Hiippala, T. (2017). *Multimodality: Foundations, Research and Analysis – A Problem-Oriented Introduction*. Berlin: De Gruyter Mouton.
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Bülow, L., Merten, M. L., and Johann, M. (2018). Internet-Memes Als Zugang Zu Multimodalen Konstruktionen. *Zeitschrift für Angewandte Linguistik* 69, 1–32. doi: 10.1515/zfal-2018-0015
- Bybee, J. (2006). From usage to grammar: the mind's response to repetition. *Language* 82, 711–733. doi: 10.1353/lan.2006.0186
- Bybee, J. (2013). "Usage-based theory and exemplar representations of constructions," in *The Oxford Handbook of Construction Grammar*, eds. T. Hoffmann and G. Trousdale (Oxford: Oxford University Press), 49–69.
- Clark, H. H. (2016). Depicting as a method of communication. *Psychol. Rev.* 123, 324–347. doi: 10.1037/rev0000026
- Council for Cultural Co-operation, Education Committee, and Modern Languages Division (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- Croft, W., and Cruse, D. A. (2004). *Cognitive Linguistics*. Cambridge: Cambridge University Press.
- Dancygier, B., and Vandelandotte, L. (2017). Internet memes as multimodal constructions. *Cogn. Linguist.* 28, 565–598. doi: 10.1515/cog-2017-0074
- Diessel, H. (2023). *The Constructicon: Taxonomies and Networks*. Cambridge: Cambridge University Press.
- Dingemanse, M., Blasi, D. E., Gary, L., Christiansen, M. H., and Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends Cogn. Sci.* 19, 603–615. doi: 10.1016/j.tics.2015.07.013
- Divjak, D. (2019). *Frequency in Language: Memory, Attention and Learning*. Cambridge: Cambridge University Press.
- Elvira-García, W. (2019). "Two constructions, one syntactic form: perceptual prosodic differences between elliptical and independent clauses in Spanish," in *Insubordination. Theoretical and Empirical Issues*, eds. K. Beijering, G. Kaltenböck, and M. S. Sansiñena (Berlin/Boston, MA: De Gruyter Mouton), 240–264.
- Féry, C. (2017). *Intonation and Prosodic Structure*. Cambridge: Cambridge University Press.
- Feyaerts, K., Brône, G., and Oben, B. (2017). "Multimodality in interaction," in *The Cambridge Handbook of Cognitive Linguistics*, ed. B. Dancygier (Cambridge: Cambridge University Press), 135–156.
- Fillmore, C. J., Lee-Goldman, R. R., and Rhodes, R. (2012). "The FrameNet construction," in *Sign-Based Construction Grammar*, eds. H. C. Boas and I. A. Sag (Stanford: CSLI), 309–379.
- Fry, D. B. (1955). Duration intensity as physical correlates of linguistic stress. *J. Acoust. Soc. Am.* 32, 765–769. doi: 10.1121/1.1908022
- Fry, D. B. (1958). Experiments in the perception of stress. *Lang. Speech* 1, 126–152. doi: 10.1177/002383095800100207
- Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago, IL: University of Chicago Press.
- Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalizations in Language*. Oxford: Oxford University Press.
- Gras, P., and Elvira-García, W. (2021). The role of intonation in construction grammar: on prosodic constructions. *J. Pragmat.* 180, 232–247. doi: 10.1016/j.pragma.2021.05.010
- Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.
- Hedberg, N., Sosa, J. M., and Fadden, L. (2003). "The intonation of contradictions in American English," in *Prosody and Pragmatics Conference*, 1–12. Available online at: [https://www.sfu.ca/~hedberg/Preston\\_paper\\_text4.pdf](https://www.sfu.ca/~hedberg/Preston_paper_text4.pdf) (accession October 13, 2023).
- Hiippala, T. (2017). The multimodality of digital longform journalism. *Digit. Journal.* 5, 420–442. doi: 10.1080/21670811.2016.1169197
- Hilpert, M. (2019). *Construction Grammar and Its Application to English*, 2nd Edn. Edinburgh: Edinburgh University Press.
- Hoffmann, T. (2017). Multimodal constructs – multimodal constructions? The role of constructions in the working memory. *Linguist. Vanguard* 3:20160042. doi: 10.1515/lingvan-2016-0042
- Hoffmann, T. (2022). *Construction Grammar: The Structure of English*. Cambridge: Cambridge University Press.
- Hoffmann, T., and Trousdale, G. (2013). *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press.
- Hsu, H. C., Brône, G., and Feyaerts, K. (2021). When gesture "takes over": speech-embedded nonverbal depictions in multimodal interaction. *Front. Psychol.* 11:552533. doi: 10.3389/fpsyg.2020.552533
- Imai, M., and Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philos. Trans. Royal Soc. B* 369:20130298. doi: 10.1098/rstb.2013.0298
- Imo, W., and Lanwer, J. P. (2020). *Prosodie und Konstruktionsgrammatik*. Berlin: De Gruyter.
- Köhler, W. (1929). *Gestalt Psychology*. New York, NY: Liveright.
- Kress, G. (2000). Multimodality: challenges to thinking about language. *TESOL Quarterly* 34, 337–340. doi: 10.2307/3587959
- Kurumada, C., Brown, M., and Tanenhaus, M. (2012). Pragmatic interpretation of contrastive prosody: it looks like speech adaptation. *Proc. Ann. Meet. Cogn. Sci. Soc.* 34, 647–652.
- Ladewig, S. (2020). *Integrating Gestures: The Dimension of Multimodality in Cognitive Grammar*. Berlin: De Gruyter.
- Lanwer, J. P. (2020). "Appositive syntax oder appositive prosodie?" in *Prosodie und Konstruktionsgrammatik*, eds. W. Imo and J. P. Lanwer (Berlin: De Gruyter), 233–281.

- Lehmann, C. (2023). "Multimodal markers of irony in televised discourse: a corpus-based approach," in *Multimodal Im/politeness: Signed, Spoken, Written*, eds. L. Brown, I. Hübscher, and A. H. Jucker (Amsterdam: Benjamins), 251–272.
- Lehmann, C. (in press). "The prosody of irony is diverse and sometimes construction-specific," in *Interfaces of Phonetics*, ed. M. Schlechtweg (Berlin: De Gruyter).
- Leiner, D. J. (2021). *SoSci Survey*. Available online at: <https://www.sosicisurvey.de> (accessed June 18, 2023).
- Lelandais, M., and Ferré, G. (2019). The verbal, vocal, and gestural expression of (in)dependency in two types of subordinate constructions. *J. Corpora Discour. Stud.* 2, 117–143. doi: 10.18573/jcads.4
- Levinson, S. C. (2006). "Deixis," in *The Handbook of Pragmatics*, eds. L. R. Horn and G. Ward (Hoboken: Wiley), 97–121.
- Levis, J. M., and Wichmann, A. (2015). "English intonation - form and meaning," in *The Handbook of English Pronunciation*, eds. M. Reed and J. M. Levis (Chichester: Wiley-Blackwell), 139–155.
- Lieberman, M., and Sag, I. A. (1974). "Prosodic form and discourse function," in *Papers from the Tenth Regional Meeting Chicago Linguistic Society*, eds. M. W. La Galy, R. A. Fox, and A. Bruck (Chicago, IL: Chicago Linguistic Society), 416–427.
- Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *J. Acoust. Soc. Am.* 32, 451–454. doi: 10.1121/1.1908095
- Long, J. A. (2022). *Jtools: Analysis and Presentation of Social Scientific Data*. Available online at: <https://CRAN.R-project.org/package=jtools> (accessed October 13, 2023).
- Lüdtke, D. (2023). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.15. Available online at: <https://CRAN.R-project.org/package=sjPlot>
- Marandin, J.-M. (2006). *Contours as Constructions. Constructions Special Volume 1*. doi: 10.24338/cons-448
- Neitsch, J., and Niebuhr, O. (2019). "Questions as prosodic configurations: how prosody and context shape the multiparametric acoustic nature of rhetorical questions in German," in *Proceedings of the 19th International Congress of Phonetic Sciences*, eds. S. Calhoun, P. Escudero, M. Tabain and P. Warren (Canberra, ACT: Australasian Speech Science and Technology Association), 2425–2429. Available online at: [https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS2019\\_Proceedings.pdf](https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS2019_Proceedings.pdf) (accessed October 13, 2023).
- Niebuhr, O. (2010). On the phonetics of intensifying emphasis in German. *Phonetica* 67, 170–198. doi: 10.1159/000321054
- Niebuhr, O. (2019). "Pitch accents as multiparametric configurations of prosodic features – evidence from pitch-accent specific micro-rhythms in German," in *A Sound Approach to Language Matters - in Honor of Ocke-Schwen Bohn*, eds. A. M. Nyvad, M. Hejná, A. Hojen, A. B. Jespersen, and M. H. Sorensen (Aarhus: Aarhus University Press), 321–351.
- Ningelgen, J., and Auer, P. (2017). Is there a multimodal construction based on non-deictic so in German? *Linguist. Vanguard* 3:20160051. doi: 10.1515/lingvan-2016-0051
- Nolan, F. (2021). "Intonation," in *The Handbook of English Linguistics, 2nd Edn*, eds. B. Aarts, A. McMahon, and L. Hinrichs (Chichester: Wiley), 385–405.
- Palan, S., and Schitter, C. (2018). Prolific Ac—a subject pool for online experiments. *J. Behav. Exp. Fin.* 17, 22–27. doi: 10.1016/j.jbef.2017.12.004
- Perniss, P. (2018). Why we should study multimodal language. *Front. Psychol.* 9:e01109. doi: 10.3389/fpsyg.2018.01109
- Pöldvere, N., and Paradis, C. (2020). 'What and Then a little robot brings it to you?' The reactive *What-X* construction in spoken dialogue. *Engl. Lang. Linguist.* 24, 307–332. doi: 10.1017/S.1360674319000091
- R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: <https://www.r-project.org/> (accessed June 23, 2022).
- Sacks, H., Schegloff, E. A., and Jefferson, G. D. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696–735. doi: 10.1353/lan.1974.0010
- Sadat-Tehrani, N. (2010). An intonational construction. *Constructions* 3, 1–13. doi: 10.24338/cons-451
- Schoonjans, S. (2018). *Modalpartikeln als Multimodale Konstruktionen: Eine Korpusbasierte Kookkurrenzanalyse von Modalpartikeln und Gestik Im Deutschen*. Berlin: De Gruyter.
- Sidhu, D. M., Westbury, C., Hollis, G., and Pexman, P. M. (2021). Sound symbolism shapes the English language: the Maluma/takete effect in English nouns. *Psychon. Bull. Rev.* 28, 1390–1398. doi: 10.3758/s13423-021-01883-3
- Singer, N. I. (2023). *Oxford English Dictionary*. Oxford: Oxford University Press.
- Steen, F., and Turner, M. B. (2013). "Multimodal construction grammar," in *Language and the Creative Mind*, eds. M. Borkent, B. Dancygier, and J. Hinnell (Stanford: CSLI), 255–274.
- Stöckl, H. (2020). "Linguistic multimodality – multimodal linguistics: a state-of-the-art sketch," in *Multimodality: Disciplinary Thoughts and the Challenge of Diversity*, eds. J. Wildfeuer, J. Pflaeging, J. Bateman, O. Seizov, and C. I. Tseng (Berlin: DeGruyter), 41–68.
- Tell, V. (2023). In *Oxford English Dictionary*. Oxford: Oxford University Press.
- Tench, P. (1996). Intonation and the differentiation of syntactic patterns in English and German. *Int. J. Appl. Linguist.* 6, 223–256. doi: 10.1111/j.1473-4192.1996.tb00096.x
- Uhrig, P. (2020). Multimodality in language and communication. *Zeitschrift für Anglistik und Amerikanistik* 68:4. doi: 10.1515/zaa-2020-2019
- Uhrig, P. (2022). Hand gestures with verbs of throwing: collostructions, style and Metaphor. *Yearb. German Cogn. Linguist. Assoc.* 10, 99–120. doi: 10.1515/gcla-2022-0006
- van Leeuwen, T. (2014). "Critical discourse analysis and multimodality," in *Contemporary Critical Discourse Studies*, eds. C. Hart and P. Cap (London: Bloomsbury), 281–296.
- Vigliocco, G., Perniss, P., and Vinson, D. (2014). Language as a Multimodal phenomenon: implications for language learning, processing and evolution. *Philos. Trans. Royal Soc. Lond. Ser. B Biol. Sci.* 369:20130292. doi: 10.1098/rstb.2013.0292
- Ward, N. G. (2019). *The Prosodic Patterns of English Conversation*. Cambridge: Cambridge University Press.
- Wells, J. C. (2006). *English Intonation: An Introduction*. Cambridge: Cambridge University Press.
- Wichmann, A., and Blakemore, D. (2006). The prosody-pragmatics interface. *J. Pragmat.* 38, 1537–1541. doi: 10.1016/j.pragma.2006.02.009
- Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., et al. (2023). *Ggplot2, Create Elegant Data Visualisations Using the Grammar of Graphics*. Available online at: <https://CRAN.R-project.org/package=ggplot2> (accessed November 13, 2023).
- Ziem, A. (2017). Do we really need a multimodal construction grammar? *Linguist. Vanguard* 3:20160095. doi: 10.1515/lingvan-2016-0095
- Zima, E. (2017). On the multimodality of [all the way from X PREP Y]. *Linguist. Vanguard* 3:20160055. doi: 10.1515/lingvan-2016-0055
- Zima, E., and Bergs, A. (2017). Towards a multimodal construction grammar. *Linguist. Vanguard* 3 :20161006. doi: 10.1515/lingvan-2016-1006





## OPEN ACCESS

## EDITED BY

Claudia Lehmann,  
University of Potsdam, Germany

## REVIEWED BY

Dimitris Serafis,  
University of Groningen, Netherlands

## \*CORRESPONDENCE

Hartmut Stöckl  
✉ hartmut.stoeckl@plus.ac.at

RECEIVED 05 January 2024

ACCEPTED 07 February 2024

PUBLISHED 22 February 2024

## CITATION

Stöckl H (2024) Fresh perspectives on  
multimodal argument reconstruction.  
*Front. Commun.* 9:1366182.  
doi: 10.3389/fcomm.2024.1366182

## COPYRIGHT

© 2024 Stöckl. This is an open-access article  
distributed under the terms of the [Creative  
Commons Attribution License \(CC BY\)](#). The  
use, distribution or reproduction in other  
forums is permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original publication in  
this journal is cited, in accordance with  
accepted academic practice. No use,  
distribution or reproduction is permitted  
which does not comply with these terms.

# Fresh perspectives on multimodal argument reconstruction

Hartmut Stöckl\*

Department of English and American Studies/English and Applied Linguistics, University of Salzburg,  
Salzburg, Austria

The present contribution exemplifies current models for argument reconstruction on an environmental protection print-ad, identifying deficits in the way the models account for multimodal argumentation. Based on this critical review, three general research perspectives are suggested for making argument reconstruction maximally multimodal: the reach and logic of semiotic modes, multimodal coherence, and genre-specific multimodal discourse structure.

## KEYWORDS

multimodal argumentation, multimodal coherence, reach of mode, argument reconstruction, discourse semantics

## 1 Introduction

Recently, the claim has been made that multimodality, rather than being an independent field of study, is “a stage of development through which many disciplines naturally pass” (Bateman, 2022, p. 41). Argumentation studies serve as a case in point, which have recognized and intensively studied multimodally expressed arguments ever since they accepted visual arguments (Birdsell and Groarke, 2007; Kjeldsen, 2015a). Multimodal argumentation has been aptly defined by Tseronis as “a communicative activity, in which more than one mode (besides spoken and written language) play a role in the procedure of testing the acceptability of a standpoint” (Tseronis, 2018, p. 12). Following Bateman’s dictum that “more needs to be done (...) than simply assuming that multimodal argumentation exists” (Bateman, 2018, p. 295), I will in this contribution critically review and exemplify selected approaches to argument reconstruction (see van Eemeren et al., 2014) for their suitability to describe the structure and functioning of multimodal argumentation, suggesting ways of enhancing the multimodal analysis. My perspective is that of a discourse linguist, who seeks to determine which place images occupy in a genre-specific multimodal argumentation and how they help constitute an argument.

## 2 Current models for argument reconstruction

### 2.1 Formal logic

Formal logic (Smith, 2007) aims to distinguish the elements in a deductive argument, which is made up of two premises and a conclusion, forming what is known as a syllogism. In the Surfrider ad (see Figure 1), the following syllogistic form may be discerned:

Premise 1:	If plastic pollution harms humans/the environment, it should be stopped.
Premise 2:	Plastic pollution harms the body as much as the ocean.
Claim (Incitive):	Say no to plastic.

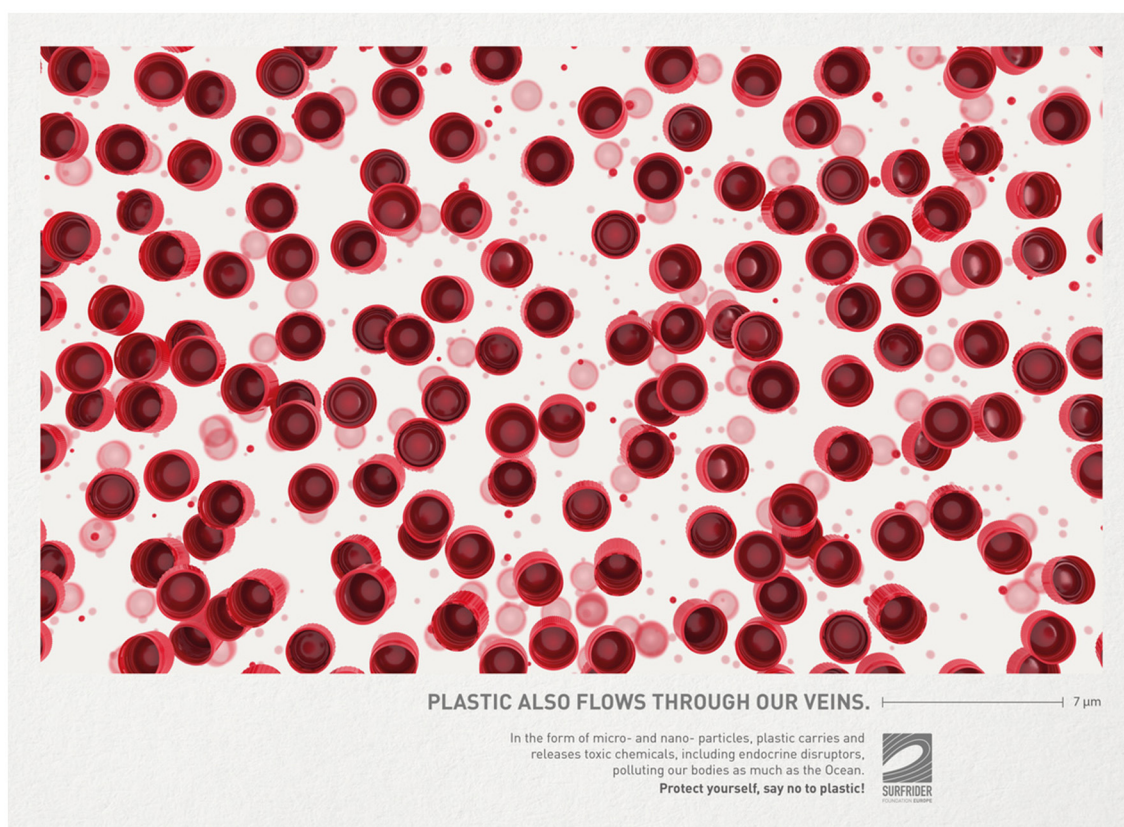


FIGURE 1  
Surfrider Foundation, France, Babel, Paris (Lürzer's Archive, 2/2022, p. 121).

This truth-conditional approach has been criticized for its artificiality. In van Rees' words: "there is a large gap between ordinary-language discourse and formal-language logic" (Van Rees, 2001, p. 179), a gap that widens considerably when we include visual/multimodal means of argumentation. In the example, the composite doctored image, which likens plastic bottle tops to red blood cells, helps express the second premise. Groarke (2015) has used the elements that establish the logical form of an argument in tables showing its key components, and demonstrates that visuals may be located there. While logical form is a methodological basis in argument reconstruction, it leaves the actual discourse context unaccounted for, most notably all knowledge of the genre.

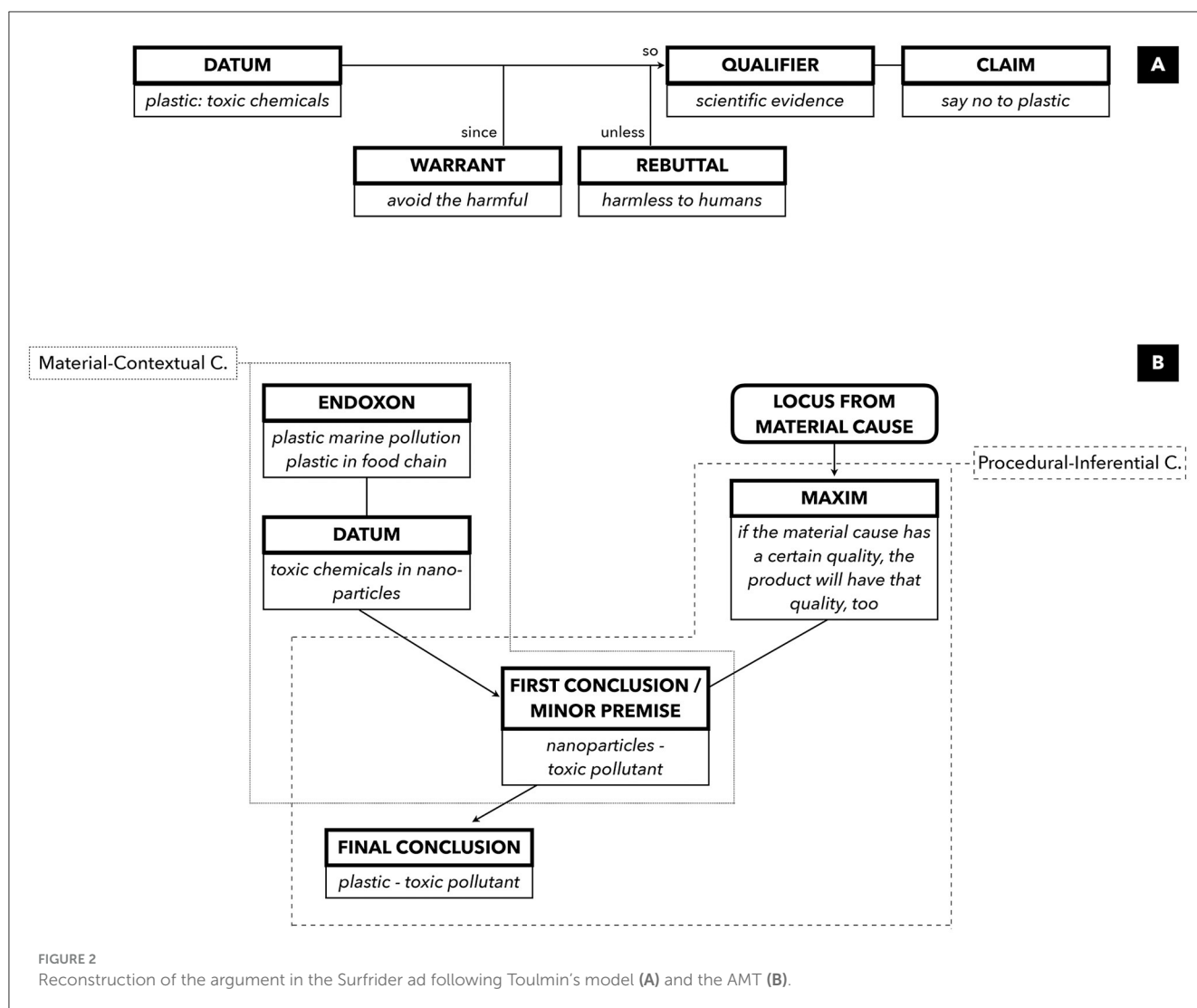
## 2.2 Toulmin's model

Toulmin's well-known model for reconstructing argument structure (Toulmin, 2008/1958, see Figure 2) essentially links a claim with data, i.e., reasons, evidence or arguments for justifying the claim. In the Surfrider ad, the *toxic chemicals, including endocrine disruptors* act as evidential data for the descriptive claim that *plastic also flows through our veins*. In turn, this claim becomes a ground to *protect yourself* and *say no to plastic*. The connection between claim and data lies in an inferential rule or principle, which Toulmin calls warrant. For the incitive claim of the ad, the warrant may be something like "if something is harmful, it

must be prevented". A fourth ingredient in Toulmin's argument structure is called qualifier and allows us to judge how reliable or valid the link between claim and data is. The text of the ad phrases the connection between plastic pollution and bodily harm as a general rule backed by science and the authority of environmental protection campaigns. However, the image with its computer-generated visual analogy between plastic particles and blood cells may give the viewer ground for doubt. Groarke (2009) suggests that visual images or visual structure can in principle (help) express all parts in Toulmin's model for argument reconstruction (see also Kjeldsen, 2012). In the sample ad, the image evidently functions as data, proving the connection between plastic and blood.

## 2.3 Pragma-dialectics

The pragma-dialectical approach (van Eemeren, 2018) views argumentation primarily as an exchange of speech acts, which become moves in a critical discussion whose participants seek to test the acceptability of a standpoint. Advertisements appear to be atypical representatives of such a critical discussion, as the genre lacks dialogic interaction and an exchange of opinion. However, this does not disqualify the pragma-dialectical approach, since we can conveniently look at an advertisement as realizing a number of argumentative moves. In our example, these are:



1. Protect yourself, say no to plastic
- 1.1 Plastic in the form of micro-/nano-particles pollutes our bodies as much as the Ocean
  - (1.1' We do not want to pollute our bodies or the Ocean)
  - 1.1.1a Plastic also flows through human veins
  - 1.1.1b Plastic carries and releases chemicals, including endocrine disruptors
  - (1.1.1a-1.1.1b' Plastic flowing through human veins and releasing chemicals is a sign that it can pollute our bodies as much as the Ocean)

By comparison with a logical approach, the examination of multimodal arguments from a speech-act perspective evidently allows us to be more explicit and to determine how individual moves are semiotically realized (see Tseronis, 2017). We can now identify moves of an argument that are made through pictures or graphics, such as 1.1 and 1.1.1a, both of which semiotically materialize as combinations of language and image. The pragma-dialectical approach has also sensitized argument analysts to premises that are left implicit and maintains that rhetors must be held responsible for such implied premises. In the advert, one

proposition is merely presupposed, namely that plastic particles really find their way into the blood stream. The visual image goes some way toward creating evidence for this proposition, but it cannot count as actual proof. Finally, pragma-dialectics has paid much attention to the inferential link between standpoint and argument(s), distinguishing three major types of argument schemes: causal, comparative and symptomatic. The Surfrider ad develops a dual causal argument: Because plastic flows through our veins, it pollutes the body and because plastic is thus harmful to humans, we must not use it. Interestingly, the visual image also implies a comparative argument scheme, i.e., plastic particles are compared to red blood cells.

## 2.4 The argumentum model of topics

The Argumentum Model of Topics (AMT) (Rigotti and Greco, 2019) ostensibly fuses a logical with a pragmatic reconstruction of argument. For this purpose, it distinguishes between two interlocking components of argument construction, a material-contextual and a procedural-inferential one (see Figure 2).

The material-contextual component is comprised of endoxa, i.e., generally accepted knowledge/opinion that is expressed, presupposed or implied in the discourse, and datum/data, i.e., facts, reasons, evidences accumulated in the discourse to support the proposed argument. In the Surfrider ad, some of the endoxical knowledge is explicit, such as knowing about ocean plastic as an environmental problem. Other endoxa are left implicit, such as the argumentatively vital knowledge about plastic in the food chain, which subsequently enters human bodies through seafood. The data brought forward are essentially about the toxic chemicals in the micro- and nano-particles that are released into the blood. The visual image contributes to expressing the datum of the argument as it literally locates micro-plastics in the molecular structure of blood. Taken together, endoxon and datum allow for a first conclusion that acts as a minor premise: “Plastic material is a toxic pollutant”. The procedural-inferential component of the argument structure combines a locus, i.e., “an ontological relation on which a given argument is based” (Rigotti and Greco, 2019, p. 210), with a maxim, i.e., an inferential rule operating on the locus. The causal locus from material cause fits the argument in the Surfrider ad best, which brings plastic (products) and nano-particles/toxic chemicals into an ontological relation. This may then be expressed as an inferential rule: “If the material cause has a certain quality, the product will have that quality, too” (Rigotti and Greco, 2019, p. 258). In an integrational synthesis, endoxon cum datum and locus cum maxim facilitate the final conclusion, i.e., the standpoint expressed in the ad: Plastic pollutes our bodies, and by implication, the advice to boycott plastic. The AMT has been used to reconstruct multimodal arguments in e.g., Serafis (2022).

## 2.5 Multimodal rhetoric

While the models exemplified so far generally allow for locating semiotic modes in argument structure, they do not specifically attend to the discourse semantics of the modes and to the ways in which they impact on the construction of the argument. Rocci et al. (2018) propose a rhetorically minded multi-layer model which inventories the different modal components of a message and inspects them for how they configure in the overall argument. Most importantly, the model assumes that verbal and visual discourse structures combine to constitute a multimodal rhetorical figure, such as metonymy or metaphor etc. In order to describe the nature of the rhetorical operation, the authors borrow the notions of “visual structure” and “meaning operation” from visual rhetoric (see Phillips and McQuarrie, 2004). In the Surfrider image, the larger plastic particles (i.e., bottle tops) are “juxtaposed” with the smaller blood cells, their identical round shapes and red colors suggesting a “comparison” and an associative “connection”. The phrases *plastic also flows through our veins* and *polluting our bodies as much as the ocean* help construe both the formal analogy and the functional association. If, as the image suggests, plastic can get into the bloodstream, this negative consequence of plastic pollution must be avoided at all cost. Such interpretations do not sideline visual images as merely “expressive” or “embellishing” add-ons (Grancea, 2017, p. 18, 21), but regard visual or multimodal rhetorical operations to be inherent facilitators of

argumentation. In this view, visual rhetorical qualities, such as presence (evidence), realism and immediacy, or semantic condensation (Kjeldsen, 2012, p. 243–244) are constitutive of multimodal argument.

## 3 Multimodal perspectives

My brief review shows that approaches to argument reconstruction have difficulties capturing the multimodal qualities of argumentation. The models do not specifically address the semiotic nature and the exact discourse contributions of the modes. Instead, the main emphasis is placed on the logical and inferential structures of the argument. Below I propose some requirements for improving multimodal argument reconstruction.

### 3.1 Modal reach and logic

First, the various modes have different “reaches” (Kress, 2010, p. 83), i.e., strengths and weaknesses for meaning making. While language/text is capable of expressing the whole spectrum of logical relations, images confront serious limitations in this regard. The visual image, on the other hand is a powerful means to display the physical properties of objects in rich detail, something referred to as “thick representation” (see Kjeldsen, 2015b). It is, therefore, plausible that multimodal arguments favor unequal mode-status relations (see Stöckl, 2020, p. 190–195), where the image is subordinated to or integrated into the discourse structure of the text. The communicative potential of an image that can be harnessed in a multimodal argument is also determined by its configuration of visual image elements and its representational style. In our example, the multiple repetition of the circular objects in various sizes and shades of red suggest a sense of “floating” in a stream. Following Kress and van Leeuwen (1996, p. 89), this is a conceptual image presenting an “analytical process”. The image is also clearly not a photographic representation of either the blood stream or of floating bottle tops. Its computer-generated qualities are vital when we consider treating the image as direct proof or evidence of the argument. Scrutinizing an image for its material-technological qualities and for its semiotic structure is an important step to a detailed description of its potential semantic contribution to a multimodal argument.

### 3.2 Multimodal coherence

Second, the hallmark of multimodal discourse is “the linking of semiotic modes and their formal, semantic and functional integration” (Stöckl, 2019, p. 53). If we determine the place of an image in the (logical) structure of an argument, something most models afford, we mainly address the functional integration of modes. An interest in formal integration would require a consideration of the layout of a multimodal text: how much space does the image occupy relative to the text? Does the image precede or follow the text, or do they alternate? Are there visual-graphic components other than the image, for example a brand logo? What about the typography (size, type, color) of the text? These and other



questions will provide relevant clues to the special multimodal linking at work in the material. The layout in the Surfrider ad makes the image a dominant entry point for the overall message, whose proximity to the headline suggests a binary unit of a verbal descriptive claim plus an image, which may either render the claim in pictorial form or add visual data. The legend-like line indicating units of size ( $7\ \mu\text{m}$ ) is a separate graphic element that relates to the image, suggesting a heavily magnified depiction, and it links to the verbal expression *micro- and nano particles*. The spatial proximity of the logo and the bolding of the incitive claim establish another formal unit, this time marking the rhetor and its call for action. Finally, semantic integration is concerned with how the modes construe multimodal coherence, i.e., a sense-continuity across modes and an inter-connectedness of elements from both modes in the form of cohesive ties (see Stöckl and Pflaeging, 2022). Such a cohesive tie is present in the Surfrider ad, where the image evokes the concept of blood and its particles floating in a stream, which relates to the words *veins/bodies* through meronymy/metonymy. The visual evocation of blood as a carrier of plastic concretizes the claim in the argument and makes the intake of plastic through food a tangible implication. Rather than take the image as a visual restatement of the claim, it is useful to think of the text-image relation as a relational proposition (see Rhetorical Structure Theory, Taboada and Mann, 2006), where the image elaborates the text through specification or illustration, and vice versa.

### 3.3 Multimodal discourse semantics and structure

Third, “arguments normally rely on an understanding of their contexts (...) in order to be meaningful” (Blair, 2015, p. 218–219). While text-internally, the various modes participating in argumentation-building provide mutual context for each another, text-externally, the single most important contextual factor is genre. It comprises knowledge about the rhetorical situation, the discourse functions, the conventional structure(s) and the appropriate semiotic style in a given discourse type. Environmental protection print-ads, for instance, typically involve such subtopics as causer, affected, problem, solution, consequences and evidence. These may be expressed in text and/or image, producing a multimodal discourse structure. In the Surfrider ad, the image shows the causer (plastic) and the affected (blood/veins/body) of pollution, whereas the text specifies these and calls upon the recipient to act accordingly. Just as genre is likely to constrain multimodal argument structure and argumentation schemes, it also determines the kinds of visuals we are likely to encounter as well as how these will be understood. In environmental protection ads, for example, denotational images may be used as truthful, indexical evidence of the harmful consequences of environmental degradation. But as our example shows, the discourse may equally well utilize CGI-images that involve quite some degree of referential fiction. The latter type of image makes visual sign configurations available that can loosely be integrated into a propositional relation with textual elements. Situating argumentation in a specific genre will also allow the analyst to determine the stereotyped propositional

content that forms the substance of the argument structure. In anti-plastic advertising, for example, causal arguments often involve marine plastics causing habitat damage and its concomitant effects on animals and humans (see Figure 1). So, rather than be content with gleaning abstract argumentation schemes, such as argument from cause or analogy, an approach centering on genre will be capable of inventorying the concrete propositions that are used in the argumentation.

## 4 Discussion and conclusion

I hope to have shown that, despite recent efforts (see e.g., Serafis and Tseronis, 2023), current models for argument reconstruction insufficiently account for the specific contributions modes other than language make to a multimodal argument. The main reason for this deficit appears to be a heavy focus on the logical structure of arguments and a neglect of the diverse ways in which non-/and para-verbal modes come to interact and cohere with the text. While van Eemeren and Grootendorst (1992, p. 64) suggest a *logical minimum* and a *pragmatic optimum* in argument reconstruction, what is required for mode-sensitive reconstructions is a *multimodal maximum*.

As I suggested, locating an image, for example, in the logical-inferential structure of an argument is a plausible start to modeling multimodal argumentation. Such an approach will of course be complicated by the fact that visual propositions do not simply act as either, standpoint, datum, or endoxa, but often help express these in indirect, implicit and covert ways. The idea that images possess a persuasive rhetorical force by providing a visual structure and a meaning operation that semantically connect to the text is another helpful step toward reconstructing the multimodal nature of argumentation schemes.

Here, I have suggested three main trajectories for future work on multimodal argumentation. First, I advocated due attention to the pragma-semantic reaches and the internal logic of a semiotic mode. This makes the analyst aware of the typical and variable properties that a mode brings to the division of semiotic labor in a process of multimodal argumentation. Second, I proposed to look in detail at how the modes combine, interact, and co-create a coherent argumentative message. This will sensitize the analysis to varying degrees and types of mode-connectedness and information-interplay. Third, I made a plea for studying multimodal argumentation not through logical abstraction but in close relation to a concrete genre with its pre-defined discourse structure. This will give the argument reconstruction the necessary contextual specificity and yield the genre-typical propositional substance of the argument.

In conclusion, “viewing problems (such as argument reconstruction—H.S.) simultaneously from contrasting disciplinary perspectives is (...) a valuable skill to be learnt” (Bateman, 2022, p. 59). The skillset required for multimodal argument reconstruction can only emerge in a productive cooperation between argumentation and multimodality researchers. An issue to be addressed in this field is a beneficial balance between discursive case-study approaches and more empirical, corpus-based approaches to multimodal argumentation (see Bateman, 2022, p. 42–43, 52–53).



## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

HS: Writing—original draft, Writing—review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the University of Salzburg Publication Fund.

## References

- Bateman, J. A. (2018). Position paper on argument and multimodality. *Int. Rev. Pragmat.* 10, 294–308. doi: 10.1163/18773109-01002008
- Bateman, J. A. (2022). Multimodality, where next? Some meta-methodological considerations. *Multim. Soc.* 2, 41–63. doi: 10.1177/26349795211073043
- Birdsell, D.S., and Groarke, L. (2007). Outlines of a theory of visual argument. *Argument. Advoc.* 43, 103–113. doi: 10.1080/00028533.2007.11821666
- Blair, J. A. (2015). Probative norms for multimodal visual arguments. *Argumentation* 29, 217–233. doi: 10.1007/s10503-014-9333-3
- Grancea, I. (2017). Types of visual arguments. *Argumentum* 15, 16–34.
- Groarke, L. (2009). “Five theses on Toulmin and visual argument,” in *Pondering on Problems of Argumentation: Twenty Essays on Theoretical Issues*, eds. F. H. van Eemeren, and B. Garssen (Dordrecht: Springer), 229–239.
- Groarke, L. (2015). Going multimodal: what is a mode of arguing and why does it matter? *Argumentation* 29, 133–155. doi: 10.1007/s10503-014-9336-0
- Kjeldsen, J. E. (2012). “Pictorial argumentation in advertising: Visual tropes and figures as a way of creating visual argumentation,” in *Topical Themes in Argumentation Theory*, eds. F. H. van Eemeren, and B. Garssen (Dordrecht: Springer), 239–255. doi: 10.1007/978-94-007-4041-9\_16
- Kjeldsen, J. E. (2015a). The study of visual and multimodal argumentation. *Argumentation* 29, 115–132. doi: 10.1007/s10503-015-9348-4
- Kjeldsen, J. E. (2015b). The rhetoric of thick representation: how pictures render the importance and strength of an argument salient. *Argumentation* 29, 197–215. doi: 10.1007/s10503-014-9342-2
- Kress, G. (2010). *Multimodality: A Social Semiotic Approach to Contemporary Communication*. London: Routledge.
- Kress, G., and van Leeuwen, T. (1996). *Reading Images: The Grammar of Visual Design*. London: Routledge.
- Lürzer's Archive. Vol. 2/2022. London: Lürzer International Limited.
- Phillips, B., and McQuarrie, E. F. (2004). Beyond visual metaphor: a new typology of visual rhetoric in advertising. *Market. Theor.* 4, 113–136. doi: 10.1177/1470593104044089
- Rigotti, E., and Greco, S. (2019). *Inference in Argumentation: A Topics-Based Approach to Argument Schemes*. Cham: Springer.
- Rocci, A., Mazzali-Lurati, S., and Pollaroli, C. (2018). The argumentative and rhetorical function of multimodal metonymy. *Semiotica* 220, 123–153. doi: 10.1515/sem-2015-0152
- Serafis, D. (2022). Unveiling the rationale of soft hate speech in multimodal artefacts: A critical framework. *J. Lang. Discrim.* 6, 321–346. doi: 10.1558/jld.22363
- Serafis, D., and Tseronis, A. (2023). The front page as a canvas for multimodal argumentation: Brexit in the Greek press. *Front. Commun.* 8, 1–14. doi: 10.3389/fcomm.2023.1230632
- Smith, V. J. (2007). Aristotle's classical enthymeme and the visual argumentation of the twenty-first century. *Argument. Advoc.* 43, 114–123. doi: 10.1080/00028533.2007.11821667
- Stöckl, H. (2019). “Linguistic multimodality – Multimodal linguistics: A state-of-the-art sketch,” in *Multimodality: Disciplinary Thoughts and the Challenge of Diversity*, eds. J. Wildfeuer, J. Pflaeging, J. Bateman, O. Seizov, and C. Tseng (Berlin/Boston: de Gruyter), 41–68.
- Stöckl, H. (2020). “Multimodality and mediality in an image-centric semiosphere: A rationale,” in *Visualizing Digital Discourse: Interactional, Institutional and Ideological Perspectives*, eds. C. Thurlow, C. Dürscheid, and F. Diémoz (Berlin/Boston: de Gruyter), 189–202.
- Stöckl, H., and Pflaeging, J. (2022). Multimodal coherence revisited: Notes on the move from theory to data in annotating print advertisements. *Front. Commun.* 7, 1–17. doi: 10.3389/fcomm.2022.900994
- Taboada, M., and Mann, W.C. (2006). Rhetorical structure theory: looking back and moving ahead. *Discourse Stud.* 8, 423–459. doi: 10.1177/1461445606061881
- Toulmin, S. E. (2008/1958). *The Uses of Argument (updated ed.)*. Cambridge: Cambridge University Press.
- Tseronis, A. (2017). “Analysing multimodal argumentation within the pragma-dialectical framework: Strategic maneuvering in the front covers of *The Economist*,” in *Contextualizing Pragma-Dialectics*, eds. F. H. van Eemeren, and W. Peng (Amsterdam: John Benjamins), 335–359.
- Tseronis, A. (2018). Multimodal argumentation: beyond the verbal/visual divide. *Semiotica* 220, 41–67. doi: 10.1515/sem-2015-0144
- van Eemeren, F. H. (2018). *Argumentation Theory: a Pragma-Dialectical Perspective*. Cham: Springer.
- van Eemeren, F. H., Garssen, B., Krabbe, E. C. W., Snoeck Henkemans, A. F., Verheij, B., and Wagemans, J. H. M. (2014). *Handbook of Argumentation Theory*. Dordrecht: Springer.
- van Eemeren, F. H., and Grootendorst, R. (1992). *Argumentation, Communication and Fallacies: A Pragma-Dialectical Perspective*. Hillsdale, NJ: Lawrence Erlbaum.
- Van Rees, M. A. (2001). “Argument interpretation and reconstruction,” in *Crucial Concepts in Argumentation Theory*, ed. F. H. van Eemeren (Amsterdam: Amsterdam University Press), 165–200.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



## OPEN ACCESS

## EDITED BY

Janina Wildfeuer,  
University of Groningen, Netherlands

## REVIEWED BY

Natalia Laba,  
The University of Sydney, Australia  
Dezheng Feng,  
Hong Kong Polytechnic University,  
Hong Kong SAR, China

## \*CORRESPONDENCE

Elena Mattei  
✉ elena.mattei@unive.it

RECEIVED 13 December 2023

ACCEPTED 12 February 2024

PUBLISHED 04 March 2024

## CITATION

Mattei E (2024) Approaching tourism communication with empirical multimodality: exploratory analysis of Instagram and website photography through data-driven labeling. *Front. Commun.* 9:1355406. doi: 10.3389/fcomm.2024.1355406

## COPYRIGHT

© 2024 Mattei. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Approaching tourism communication with empirical multimodality: exploratory analysis of Instagram and website photography through data-driven labeling

Elena Mattei\*

Department of Linguistics and Comparative Cultural Studies, Ca' Foscari University of Venice, Venice, Italy

This paper reports the methods and results of the manual annotation of visual features in two corpora of tourism photography on travel boards' digital channels with a tailored tagging model based on the Grammar of Visual Design and adapted to tourism discourse. Computational analysis and statistical modeling show how the testing of theoretical assumptions through categorized data may lead to evidence-based interpretations of patterns of data clustering and to the detection of new communicative aims and conventions across digital media. Preliminary findings reveal indeed significant differences in the frequency of tag (co)patternings and use of visual strategies across channels that are related to the role and aim of each channel in the marketing funnel of persuasion and journey toward purchase (AIDA). Instagram imagery was demonstrated to foster a pre-consumption of the travel experience and emotionally charged reactions by representing perceptive and emotive expectations. While both channels play on postmodern tourists' desire for the uncontaminated, remote and the authentic, Instagram favors aerial views of pristine, aesthetically pleasant settings, often complemented with rear views of solitary individuals performing static processes of contemplation of natural wonders. This suggests a focus on attracting the attention and providing instant gratification of the senses by representing what stands in contrast to everyday life and traditional tourist experiences, both avoiding cognitive effort in a pervasive digital sphere with endless sources of information and encouraging further exploration on websites.

## KEYWORDS

empirical multimodality, tourism discourse, visual communication, Instagram photography, annotation system, exploratory statistics, data mining

## 1 Introduction

Since the advent of digital platforms, photographs have been easily disseminated to shape perspectives and elicit emotions, due to their perceptual similarity to external experiences and their seemingly unmediated nature (Mitchell, 1994). Their aesthetically appealing design is argued specifically to facilitate the "rapid delivery, ubiquitous availability and the instant gratification of desires" in a rapidly changing world (Tomlinson, 2007, p. 74).

Tourism narratives, particularly, have been harnessed to promote a nostalgic, romantic view of travel destinations that transcends mass-targeted perspectives, as they play on socially driven desires of postmodern individuals and encourage the latter to find satisfaction by seeking an escape from over-regulated, over-socialized work routines and pre-packaged holiday experiences (Urry and Larsen, 2011; Mattei, 2023b).

In the funnel of persuasion or AIDA model—an acronym that stands for *Attention, Interest, Desire, Action*—Instagram (IG) is used<sup>1</sup> to evoke and raise awareness of particular sensations by means of tourist destination promotion, often generating favorable attitudes and purchase intentions (Sormaz and Ruoss, 2020). In the travel industry, it is estimated that 95% of the major brands own an IG account<sup>2</sup>, as social media marketing increases traffic to sales platforms like websites (Hays et al., 2013; Leung et al., 2013). Particularly, IG images have been shown to change tourists' perceptions and behaviors with aesthetically appealing content and *landscape* representations (Garrod, 2009, p. 355; Shuqair and Cragg, 2017). This may be due to the presence of visual features that instantly gratify the senses and stimulate immediate reactions, without requesting cognitive effort in a digital sphere pervaded with endless sources of information.

## 1.1 The scope of the study

To verify the specific role, aim of Instagram in tourism discourse and marketing, the nature of imagery in three popular tourist boards' Instagram and website pages was analyzed with social semiotics and empirical methods. This was done by looking at visual features, or *materialities* (Bateman, 2018), their frequency and clustering across digital channels and by testing the validity of the main framework available in multimodality research for image analysis, i.e., the linguistics-based *Grammar of Visual Design* (Kress and van Leeuwen, 2006). Such qualitative framework, indeed, is grounded in a general and presumed symbolic, socially attached value of images that may lack empirical testing and potentially objectivity (Bateman, 2019a, p. 533–535; Bateman, 2019b, p. 90–92). By studying larger amounts of data with bottom-up approaches, conversely, theories on the presence of tourist driving forces, or desires—such as the quest for the *uncontaminated, unknown, remote*—may be validated (Mattei, 2023c).

This procedure was guided by two research questions:

- Are there systematic visual choices in tourism, digital communication?
- Are there any differences in the use of the visual mode across media?

Through the development of a tagging system and the conduction of bottom-up statistical analysis for the detection of data clustering and dimensions of variation, it becomes visible

how Instagram imagery varies from website narratives, and seems to be systematically constructed to capture the attention and foster both a perceptual pre-consumption of the travel experience and emotionally charged reactions by representing visually such conditions. Indeed, Instagram mainly features aerial views of pristine, aesthetically pleasant settings, often complemented with rear views of solitary individuals performing static *processes* of peaceful contemplation of natural wonders. IG photographs are shown specifically to be taken from either long shots and high angles in natural settings, often without human footprint, or in close contact with wild animals. Thus, IG is shown to play a particular role in the marketing funnel of persuasion of tourists, slowly becoming “a historically stabilised site for the deployment and distribution of some selection of semiotic modes for the achievement of varied communicative purposes” that may depend on the channel's audience (Bateman et al., 2017, p. 123). The findings were obtained by analyzing the presence of particular participants, their actions and gaze in specific settings, as well as the shots, angles, and salience or positioning in the picture.

Website photography, on the other hand, is demonstrated to feature mostly active and human participants photographed in collective and social moments, also from close distance, and involved in a variety of gastronomic and cultural events in artificial settings like cities. The importance of providing information and agency is confirmed by the predominance of humans involved in activities or in the fruition of services.

## 2 Methods

This section describes data collection (Section 2.1), theory adaptation and data annotation (Section 2.2), and the use of SRI Tagging software for manual (inter)annotation and data export for statistical analysis (Section 2.3). This section thus illustrates how linguistics theories may be tested and tailored to suit genre and discourse-specific data and support an integrated, systematic multimodal analysis (Bateman, 2014, forthcoming).

### 2.1 Data collection and theory adaptation for data-driven annotation

For the project, three popular national tourist boards located in English-speaking countries were selected, especially for their daily sharing of high-quality images combined with long captions. Data consisting of photographs and corresponding texts were gathered from two main communication platforms:

- *Instagram*, used to enhance brand visibility and establish an emotional attachment, and
- the *company website*, the primary revenue source offering booking opportunities.

The corpora were collected in 2019 by accessing Instagram's application programming interface (API) to retrieve posts for each tourist board in a timespan of 6 months of the same year, and by scraping websites once permission was granted. Table 1 presents a

1 <https://www.statista.com/statistics/259379/social-media-platforms-used-by-marketers-worldwide/> (accessed January 29, 2024).

2 <https://www.statista.com/statistics/499694/forecast-of-online-travel-sales-worldwide/> (accessed January 29, 2024).

TABLE 1 Subdivision of visual sub-corpora according to tourist board and channel (Mattei, 2023a, p. 5).

Tourist board and sub-corpus	Number of images (from instagram posts)	N. of images (from official websites)	N. of total images per tourist board
Tourism Ireland	126	200	326
Destination Canada	180	160	340
Tourism Western Australia	178	158	336
Total	484	518	1,002

breakdown of the visual sub-corpora, categorized by tourist board and their respective channels.

2.2 Metafunctional meaning annotation: tree tagging modeling

The statistical analysis of visual strategies in tourism photography was supported by a tree tagging system based on the *Grammar of Visual Design* but adapted to tourism photography (Mai et al., 2011). The model enabled the quantification of frequencies and correlations between objective features, like the type of settings, entities, actions or reactions, shots, gazes, photography techniques that are of interest to multimodalists (Kress and van Leeuwen, 2006). This empirical process enabled to cautiously move from theory to evidence-based, sociological interpretations of the data.

This model thus tested an adapted version of a general model of visual knowledge by categorizing and measuring occurrences, variations of data-informed visual features after close data inspection and (inter)annotation. The tags were then grouped by *metafunction* (Halliday and Matthiessen, 2014).

Specifically, the model examines the decisions made by discourse specialists regarding the selection of *participants* (subjects), the presence and type of *processes* (activities), and *settings* (macro-category *Representation of Reality*); the selection of camera shots and angles (macro-category *Relationship with the Audience*), and both the position, size and visual weight of elements within the photograph (macro-category *Composition*).

For preliminary analyses, the macro-category *Tourism Strategies* was included to quantify the presence of tourists’ driving forces or recurring themes (Maci, 2020) based on the identification of particular features in each image<sup>3</sup>. For example, the search for what is unknown, remote or pristine was encoded as *strangerhood* (trope) when detecting uncontaminated environments, like exotic beaches or clear waters. Or again, the representation of meals, swimwear or alcohol was encoded as *indexical reference*, an object symbolizing postmodern desires and taboos embodied by represented participants and transferred to tourists (Dann, 1996).

3 This macro-category was subsequently excluded to avoid redundancy.

TABLE 2 Tree tagging system for the annotation of tourism images building on Halliday’s three metafunctions and the Grammar of Visual Design (Mattei, 2023a, p. 16–19).

Macro-category (grouping features that are always selected)	Sub-category (macro-variables) (grouping features that are often selected)	Main variables (choices)
Representation of reality	Participants	<ul style="list-style-type: none"><li>• Humans</li><li>• Animals</li></ul>
	Processes	<ul style="list-style-type: none"><li>• Action</li><li>• Reaction</li><li>• Transactional action/ reaction</li><li>• Non-transactional action/ reaction</li></ul>
	Setting	<ul style="list-style-type: none"><li>• Natural</li><li>• Artificial</li><li>• Cultural</li><li>• Historical</li><li>• Gastronomic</li><li>• Analytical</li></ul>
Relationship with the audience	Gaze	<ul style="list-style-type: none"><li>• Toward the represented participant</li><li>• Toward the interactive participant</li></ul>
	Camera shot	<ul style="list-style-type: none"><li>• Close</li><li>• Medium</li><li>• Long</li><li>• Very long</li></ul>
	Camera angle	Subjective image <ul style="list-style-type: none"><li>• Vertical angle (low, eye-level, high, very high)</li><li>• Horizontal angle (frontal, oblique)</li></ul> Objective image <ul style="list-style-type: none"><li>• Direct frontal</li><li>• Perpendicular top-down</li></ul>
Composition	Space distribution	Rule of thirds <ul style="list-style-type: none"><li>• One/more points</li><li>• Scenic rule of thirds (water, land, sky)</li><li>• Lines (horizontal, vertical)</li></ul> Other techniques <ul style="list-style-type: none"><li>• Centric</li><li>• Polarization</li><li>• Symmetry</li></ul>
	Visual flow	<ul style="list-style-type: none"><li>• Leading line(s)</li><li>• Connecting dots</li><li>• Framework</li></ul>
	Visual weight	<ul style="list-style-type: none"><li>• Landscape element</li><li>• Represented living participant</li><li>• Object</li></ul>

The macro-categories linked to the three linguistic metafunctions are summarized in Table 2 and encompass a range of sub-categories of tagging possibilities (or *macro-variables*) that group “choices” in meaning construction (*main variables*).

In the tagging system, the macro-category *Representation of Reality* relates to the *experiential* metafunction and classifies both *narrative* and *conceptual* images and related *circumstances* (Kress and van Leeuwen, 2006, p. 48–62)<sup>4</sup>, i.e., settings, here identified

4 The categories *Activities* and *Means* are not included here for reasons of space.

and coded by the author as natural, artificial, historical, cultural, analytical and gastronomic. In particular, narrative images are detected by coding action and reaction (observation) *processes*<sup>5</sup>, together with the subject and nature of the process, i.e., whether it is transactional or non-transactional and therefore includes or not the *goal* of the action or the object of contemplation. Conceptual images, conversely, are articulated into *analytical* (maps, icons, logos) or *symbolic suggestive*, coded implicitly through the absence of human beings.

The macro-category *Relationship with the Audience* explores choices concerning the relationships between the producer of the semiotic artifact, the represented participants, and the interpreter who makes sense of the sign in the social context, i.e., the viewer. The producer communicates with the viewer through three main tools, which correspond to the categories included: (a) the *gaze*, if present, directed toward the viewer or a represented participant; a close, medium, or long, very long *shot*, which progressively provide a bigger picture with a less focus on details, as close shots, for example, capture only the portions of an entity; and an *angle*, which may be objective (top-down or frontal, neutralizing any distortions) or subjective. The adoption of a subjective perspective includes choices regarding vertical (high, eye-level, low) and horizontal angles, the latter being oblique when the frontal plane of the photographer does not run parallel to the one where the main represented participants are.

## 2.3 Software for manual annotation and inter-coder consistency

The hierarchical tagging system was created using Statistically Reliable Image Tagging<sup>6</sup>, a software designed for complex manual tagging procedures on large visual corpora<sup>7</sup>. SRI Tagging was developed to provide a user-friendly interface for categorizing lexicogrammatical units in multimodal studies<sup>8</sup>.

To ensure the objectivity and consistency of the tagging process and rule out the possibility of chance agreement, inter-coder reliability measures such as Cohen's kappa and Krippendorff's alpha were implemented. These were further elaborated in R to consider tag dependencies and both mutually and non-exclusive child nodes (Bateman et al., 2017, p. 198–204). In this project, 18% of the images were independently tagged by another coder, instructed on tag meanings through a reading scheme. Most variables in this study report substantial reliability values (>0.61), ensuring the robustness of the tagging procedure<sup>9</sup>.

5 In the *Grammar of Visual Design*, reaction processes involve the static observation of an entity on the part of a living subject (*participant*).

6 SRIT is an openly accessible software, designed in collaboration with Pibiri (National Research Council in Italy, ISTI-CNR); Pibiri and Mattei (2020).

7 Link: <http://xor.isti.cnr.it:8000/login.html> (accessed 30 January, 2024).

8 SRIT has been used in multimodality and digital humanities courses at Verona and Bremen universities and in international workshops.

9 A list of reliability values is provided in Mattei (2023a, p. 23–24) and is partly included in the [Supplementary material](#).

## 3 Results

### 3.1 Data extraction, grouping and inferential statistics

This section presents the results derived from the manual tagging of the visual corpora through Principal Component Analysis (PCA) in R. In particular, this section offers an overview of data patterns and clustering, including the contribution of each variable to statistical variation. PCA reduces the complexity of the data under investigation and explores “to what extent the annotations allow the data to be grouped into clusters” (Bateman and Hiippala, 2021, p. 7) by means of fictitious dimensions of variance. Due to its bottom-up nature, this test enables the generation of data-driven hypotheses of variance, correlation by mapping data onto a multidimensional space. Eventually, this analysis helped understand whether tags (i.e., variables) grouped according to channel, preparing the ground for inferential statistical analyses such as One-Way ANOVA in Jamovi, chi-square and Correspondence Analysis (CA) in R within and across macro-categories<sup>10</sup>. These tests helped identify cluster patterning and correlations, i.e., variations in the use of specific tags across different channels (Field et al., 2012). This was possible after grouping of each image's tag occurrences of each variable according to channel. To conduct ANOVA tests, both Shapiro-Wilk test of normality of the distribution of the data and Levene test were performed, reporting non-significant values and confirming the assumption of normality and homogeneity of variance<sup>11</sup>.

### 3.2 Principal component analysis: explained variance and data clustering

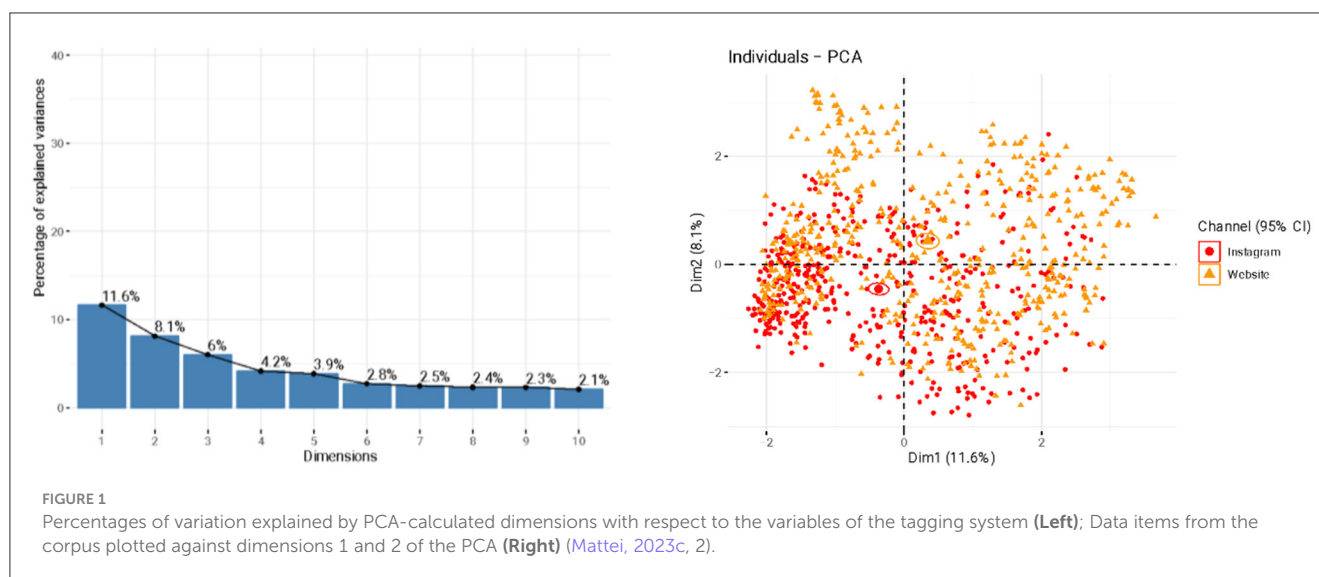
The input data used for the calculation of variation across dimensions was a table with a list of all images and the presence or absence of each feature (tag) for each image. All metadata pertaining to the type of sub-corpus was excluded and mapped afterwards. The first round of exploratory PCA includes all variables, without any filtering based on reliability values<sup>12</sup>. This was done to provide an overview of potential clustering patterns, and due to space constraints, considering the complex

10 The *chi-square* test proved to be suitable for the comparison of continuous values (number of occurrences) of nominal variables (tags) among groups, and to confirm or reject the hypothesis of equal variance of these variables among digital channels (Field et al., 2012). In this study, the Chi-square test validated the hypothesis that some features identified in the visual corpora of promotional imagery vary “unexpectedly” in terms of frequency depending on the channel of dissemination (the predictor variable) (Mattei, forthcoming).

11 The Levene test reported non-significant values for almost all variables (98%). The results of the One-Way ANOVA were successively compared with those of the Chi-square test (Mattei, 2023b, forthcoming).

12 To improve reproducibility and replicability of the results, the R script used for the statistical analysis is openly accessible at <https://github.com/ElenaMattei/Tourism-Photography>.





procedure of implementation of reliability tests on hierarchical systems. Analyses excluding non-reliable and redundant features have been extensively reported in Mattei (2023b) and partially in Mattei (2023a) and Mattei (forthcoming) and describe the implementation of Cohen's kappa and Krippendorff's alpha. To facilitate comparison with more reliable results, Supplementary material report a PCA ( $\alpha > 0.8$ ) and CA (Kappa and  $\alpha > 0.6$ ) of reliable features only.

An overview of the percentage of explained variance is provided in Figure 1, in the graph on the left side. As can be seen from the scree plot, the first ten dimensions account for 45.9% of the variance detected by the Principal Component Analysis. Each one of the remaining 191 dimensions detected accounts for too little variation of the data; this phenomenon might be investigated further by annotating a larger body of images and providing more annotations for each one of the (sub)categories of the tagging system, or by reducing the number of variables (tags) in the taxonomy. For reasons of space, the current section focuses only on the exploration of the first two dimensions of variance, which contribute the most to the explained variance.

The graph on the right side shows the mapping of each tagged image onto a multidimensional space. The PCA allowed for a preliminary observation of potential regularities across the dataset with respect to their use of visual elements according to channels. In particular, each image was plotted according to the presence or absence of the features (variables) of the tagging system, and images displaying a similar pattern of occurrences and co-occurrences of features clustered together. The specific weight, or contribution, of each variable to the first two dimensions of variance calculated allowed for a particular placement of each image by taking into account its sum of features and related scores.

As can be noticed from the graph, two distinct clusters may be distinguished, occurring in opposing quadrants along the first and second dimension. The presence of two different clusters shows how selected classes of data items group together. Clusters are represented graphically as ellipses referring to the mean of a set of values with a 95% confidence level (Bateman et al., 2021, p. 16). In particular, the spread of the ellipses along the dimensions

indicates the deviation of the data samples from the mean values. In the current graph, the narrow ellipses suggest that there is not considerable variation within the samples, i.e., website and Instagram images.

This finding is significant if accompanied by the interpretation of the position of the two distinguished clusters along dimensions 1 and 2. Indeed, website images are grouped together in the top right quadrant, showing a positive contribution to both dimensions. Conversely, Instagram images cluster in the bottom left quadrant, which indicates a negative contribution to the same dimensions. This distinction suggests that digital, visual communication in tourism discourse is constructed differently depending on the channel of dissemination of the promotional message.

In terms of variation across tourism boards, the PCA showed a less distinct and significant difference between the communicative, visual strategies; indeed, Tourism Ireland<sup>13,14</sup> and Tourism Western Australia's<sup>15,16</sup> ellipses overlap in the same quadrant, whereas only Destination Canada's<sup>17,18</sup> images cluster in the opposite quadrant, but within closer distance, compared to the graph reported (see Supplementary material).

13 Ireland (2023). Tourism Ireland. Available online at: <https://www.ireland.com/en-us/> (accessed December 13, 2023).

14 Tourism Ireland [@tourismireland] (2023). Tourism Ireland. Available online at: <https://www.instagram.com/tourismireland/> (accessed December 13, 2023).

15 Tourism Western Australia (2023). Tourism Western Australia. Available online at: <https://www.westernaustralia.com/en/welcome-to-western-australia> (accessed December 13, 2023).

16 Tourism Western Australia [@westernaustralia] (2023). Tourism Western Australia. Available online at: <https://www.instagram.com/westernaustralia/> (accessed December 13, 2023).

17 Destination Canada [@explorecanada] (2023). Destination Canada. Available online at: <https://www.instagram.com/explorecanada/> (accessed December 13, 2023).

18 Keep Exploring. (2023). Destination Canada. Available online at: <https://us-keepexploring.canada.travel/> (accessed December 13, 2023).

TABLE 3 Variables' contribution values to dimensions 1 and 2.

Variable	Contribution to dimension 1	Variable	Contribution to dimension 2
Direct presence of living beings	0.4142980	Traditional/typical	0.1526744
Subject of action	0.3590770	Artificial	0.1030312
Process of action	0.3590770	Cultural	0.1027947
Activities	0.3536642	Horizontal angle	−0.1012964
Type of action	0.3536531	Transactional reaction	−0.1015420
Humans (A)	0.3447821	Subjective image	−0.1024202
Direction of gaze	0.2903271	Vertical angle	−0.1057708
Toward the represented participant (s)	0.2671910	Oblique	−0.1059388
Visual weight	0.2195084	Converging lines—vanishing point(s)	−0.1094236
Represented living participant	0.2136750	Landscape element	−0.1129814
Transactional action	0.2094480	Direct presence of living beings	−0.1136955
Medium	0.1824133	Long	−0.1281442
Non-transactional action	0.1480952	Humans (R)	−0.1510752
Eye-level	0.1368400	Type of reaction	−0.1529681
Sport activities	0.1082395	Process of reaction	−0.1560180
Element in the foreground	0.1049749	Subject of reaction	−0.1560180
Couple	0.1031090	Rule of thirds	−0.1921793
Connecting dots	0.1008925	Water	−0.2078572
Natural	−0.1095064	Marine environment	−0.2319587
Marine environment	−0.1138014	Mountains	−0.2535981
Very long	−0.1874336	Negative space	−0.2954747
		Land	−0.3017383
		Natural	−0.3095157

### 3.3 Contributions of variables to the largest PCA-calculated dimensions of variation

The characterization of the first dimensions of variance, which account for the largest contribution to variation in the corpora under investigation, allowed to discuss the potential nature of variance in the corpus. Particularly, the calculation of the weightings of each variable sharpens an understanding of the variables behind variance in the data items, i.e., the images.

Table 3 shows the values of the variables which contribute the most to the largest dimensions of variance that were calculated by the PCA. The features with values  $> 0.1$  contribute positively to a placement along the corresponding dimension; conversely, features with weightings  $< -0.1$  contribute negatively to the dimension, indicating the presence of an inverse correlation between the dimension and the variables. For ease of reading, results report only variables which proved to be reliable in subsequent analyses.

As can be seen from the table, the variance in the first dimension is mostly due to positive contributions, whereas the second is mainly characterized by negative ones. As website images cluster along the top-right quadrant, showing positive values of significant variation in both dimensions, their nature may be said to be defined by dimension 1. The

latter includes the variables *Direct Presence of Living Being*, which encompasses animal and human life, *Process of Action* performed by *Humans (A)*, the sub-category of gaze *Toward the Represented Participant(s)* combined with *Medium* (shots), *Eye-level* (angle), and *Represented Living Participant* (a sub-category of *Visual Weight*).

Conversely, Instagram imagery, showing negative values of variation in the graph, seems to be defined by dimension 2, and in particular by the variables *Transactional Reaction*, *Oblique* (Angle) and *Vertical Angle*, *Converging Lines*, *Landscape Element* (a sub-category of *Visual Weight*), and *Direct Presence of Living Being*. These are combined with *Process of Reaction* and *Humans (R)*, *Long* (Shot), *Rule of Thirds* and correlated *Water*, *Mountains*, *Land*, which encode large and definite portions of pictures representing landscapes. If we consider the variables' presence, weight, and positive or negative value, we may formulate the following hypotheses, confirmed by the chi-square and CA:

- Dimension 1, which includes mostly positive contributions, thus representing website imagery, correlates positively with the presence of living beings performing *action* processes from a social distance (*medium* shots) and *eye-level* angles. It also correlates with the presence of *gazes*;

- Dimension 1 is negatively correlated with long shots of natural environments;
- Dimension 2, which includes mostly negative contributions, thus representing Instagram imagery, correlates negatively with natural representations of the destination as well as rear views of single beings involved in *reaction* processes from *long* shots and subjective angles, specifically *vertical* and *oblique*. The attention to the visual representation is enhanced through the use of professional perspective techniques, including converging lines.
- Dimension 2 correlates positively with traditional and cultural aspects of tourist destinations, like customs, music festivals or events, local attractions and landmarks, or objects typical of a culture or history of a population.

On the one hand, results show website communication focuses on *narrative* representations that visually shape a multifaceted tourist experience, in which humans are the main characters, occupying a good portion of the photograph, and are involved in various activities. These include sports, collective gastronomic experiences, and tours with guides or in more exclusive conditions (e.g., by car, on a cruise). Also, the presence of *gazes* and *eye-level* angles shapes a close and equal relationship with viewers (Messaris, 1997), possibly mirroring instances of everyday life or mass tourism (Urry and Larsen, 2011). The representation of activities in various settings, including content like maps, signals an attempt to show what can be practically *done* and where when visiting a new place. This is because websites are visited mostly by individuals who have developed informative or purchase intents related to a brand or service, increasingly influenced by social media's awareness- and community-building activities through content sharing and interaction with a broad audience (Leung et al., 2013; Jamil et al., 2022)<sup>19</sup>.

Conversely, Instagram communication is defined by *conceptual* and *static narrative* representations in which a distant, commodifying view of the tourist destination is constructed, often combined with professional compositional techniques like *converging lines* and *Rule of Thirds* that make the picture attractive to the human eye (Mai et al., 2011). This is line with computational studies showing how posting aesthetically appealing photography is a shared practice in the IG community (Manovich, 2017). Specifically, the combination of *long* shots of vast, natural views and rear views of *reaction processes* with *vertical angles* establishes an impersonal relationship with dehumanized subjects who, when present, rarely engage with viewers through *gazes*; rather, they perform solitary, static processes of *perceiving* natural wonders, as also reported by qualitative and marketing studies (Garrod, 2009; Smith, 2021). Through such semiotic acts, represented tourists become themselves commodities, symbolizing consumption and fostering identification (Debord, 1977; Francesconi, 2014). The use of vertical angles—often *high*, as reported by absolute counts and CA—might signal the intention to build an unequal relationship between represented participants and viewers, as the latter view the

destination from a superior position denoting control, ownership and power (Kress and van Leeuwen, 2006). Compared to website imagery, thus, Instagram imagery prompts imagination through *scapes* and a sense of pristine, remote experience, defined *extraordinary* by tourism sociologists (Urry and Larsen, 2011), i.e., different from everyday routines and mass-targeted, collective experiences in industrialized, polluted areas (Dann, 1996).

## 4 Discussion

The PCA confirmed the presence of variance in the visual construction of travel destinations across media. Overall, the study shows how images are designed to shape positive expectations about intangible leisure experiences, potentially manipulating perceptions and legitimating (pre)consumption practices.

The paper also discussed the implications of testing theories and annotating with data-driven procedures in a supervised environment, both emphasizing the exploration, adaptation of general knowledge theories and exploring multimodal frameworks' suitability for analyzing meaning-making processes in particular contexts. The importance of empirical testing on larger datasets was highlighted to avoid imposing classifications without verification. Both data and testing were key to provide evidence for the validation of previous qualitative insights concerning the role of the visual mode in constructing discourse semantics and social meaning (Stöckl et al., 2020). This brief report thus positions itself in the empirical multimodal literature and alongside quantitative content analyses (Bouko et al., 2021), and was followed by an extensive, qualitative discussion of reliable findings (Mattei, 2023b).

The annotation of a specialized corpus is also an expansion of quantitative work and an attempt to explore new communication technologies through empirical setups that substantiate hypotheses and foster comparisons with other disciplines. Indeed, the findings contribute not only to literature on language-based models and social semiotics, but also to marketing studies, which have recently addressed the social media emerging trend toward romantic visualizations of the tourist experience (Cilkin and Cizel, 2022), yet leaving institutional communication unaddressed.

The paper also highlights the role of manual annotation in enabling customization and control over data classification. Cultivating analytical skills through close inspection and elaboration of taxonomies may facilitate the interpretation, generation of signs with informed social understandings, shedding light on how subjective labeling both impacts content analysis and offers deeper insights into meaning making practices that go beyond entity recognition. Creating new taxonomies may thus become a way to train pattern recognition and identify features that are meaningful to multimodality researchers, including relationships between represented participants, settings, actions, gazes, shots, and compositional techniques. Eventually, this procedure may inform other disciplines and validate previous studies; it can also foster digital literacy in an era dominated by misinformation and AI-generated content.

Despite the corroboration of the findings through CA and Chi-square test, further model testing is required. PCA is an emerging practice in multimodal analysis for the exploration of variance that may be related to metadata. The presence of many dimensions highlights the need to annotate a larger body of images and provide

<sup>19</sup> In Mattei (2023b), website language is also shown to be more informative compared to Instagram's. The multimodal analysis allowed for an investigation of text-image relationships within and across channels (Bateman, 2014).

more annotations for each label. The training of a model that explains variance more predictively, as shown by Computer Vision studies, may prompt comparison between (un)supervised entity recognition through machine learning procedures and customized annotation systems that build on and test theoretical systems of knowledge.

Finally, the administration of surveys might contribute to existing marketing research on IG followers' intentions and changes in perception and behavior after exposure to promotional content.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

## Ethics statement

Ethical approval was not required for the study involving human data in accordance with the local legislation and institutional requirements. The social media data was accessed and analyzed using Instagram's API. The Instagram and website content were collected once permission was granted by the travel boards.

## Author contributions

EM: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Software, Validation, Writing—original draft, Writing—review & editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The

research activities presented in this paper were conducted at the University of Verona, in the Department of Foreign Languages and Literatures, as part of the Excellence Programme in Digital Humanities. The department and the Italian Ministry of University and Research funded the programme.

## Acknowledgments

The analysis reported in this paper was conducted thanks to the invaluable support and expertise of John Bateman during a research period at Universität Bremen. Any errors, omissions or misrepresentations remain solely the author's responsibility.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcomm.2024.1355406/full#supplementary-material>

## References

- Bateman, J. A. (2014). *Text and Image: A Critical Introduction to the Visual/Verbal Divide*. London: Routledge.
- Bateman, J. A. (2018). Peircean semiotics and multimodality: towards a new synthesis. *Multim. Commun.* 7, 1–24. doi: 10.1515/mc-2017-0021
- Bateman, J. A. (2019a). Towards critical multimodal discourse analysis: a response to Ledin and Machin. *Crit. Disc. Stud.* 16, 531–539. doi: 10.1080/17405904.2018.1550430
- Bateman, J. A. (2019b). "The critical role of analysis in moving from conjecture to theory," *Shifts Towards Image-Centricity in Contemporary Multimodal Practices*, eds H. Stöckl, H. Caple, and J. Pflaeging (London: Routledge), 86–94.
- Bateman, J. A., Heller, V., Moschini, I., and Sindoni, M. G. (forthcoming). "What is genre? Foundational considerations for genre and some implications for methods," *Multidisciplinary Views on Discourse Genres*, eds N. Stukker, J. Bateman, D. McNamara, and W. Spooren (London: Routledge), 1–42.
- Bateman, J. A., and Hiippala, T. (2021). "From data to patterns," in *Empirical Multimodality Research: Methods, Evaluations, Implications*, eds J. Pflaeging, J. Wildfeuer, and J. A. Bateman (Berlin: Walter de Gruyter), 65–90.
- Bateman, J. A., Veloso, F. O. D., and Lau, Y. L. (2021). On the track of visual style: a diachronic study of page composition in comics and its functional motivation. *Visual Commun.* 209–247. Available online at: <https://journals.sagepub.com/doi/abs/10.1177/1470357219839101?journalCode=vcja>
- Bateman, J. A., Wildfeuer, J., and Hiippala, T. (2017). *Multimodality: Foundations, Research and Analysis—A Problem-oriented Introduction*. Berlin: De Gruyter Mouton.
- Bouko, C., De Wilde, J., Decock, S., De Clercq, O., Manchia, V., and Garcia, D. (2021). Reactions to Brexit in images: a multimodal content analysis of shared visual content on Flickr. *Vis. Commun.* 20, 4–33. doi: 10.1177/1470357218780530
- Cilkin, R., and Cizel, B. (2022). Tourist gazes through photographs. *J. Vacat. Market.* 28, 188–210. doi: 10.1177/13567667211038955
- Dann, G. (1996). *The Language of Tourism: A Sociolinguistic Perspective*. Wallingford: CAB International.
- Debord, G. (1977). *The Society of the Spectacle*. Detroit: Black & Red.
- Field, A., Miles, J., and Field, Z. (2012). *Discovering Statistics using R*. London: Sage.

- Francesconi, S. (2014). *Reading Tourism Texts: A Multimodal Analysis*. Bristol: Channel View Publications.
- Garrod, B. (2009). Understanding the relationship between tourism destination imagery and tourist photography. *J. Travel Res.* 47, 346–368. doi: 10.1177/0047287508322785
- Halliday, M. A. K., and Matthiessen, C. M. I. M. (2014). *Introduction to Functional Grammar*. London: Routledge.
- Hays, S., Page, S. J., and Buhalis, D. (2013). Social media as a destination marketing tool: its use by national tourism organisations. *Curr. Iss. Tour.* 16, 211–239. doi: 10.1080/13683500.2012.662215
- Jamil, K., Dunnann, L., Gul, R. F., Shehzad, M. U., Gillani, S. H. M., and Awan, F. H. (2022). Role of social media marketing activities in influencing customer intentions: a perspective of a new emerging era. *Front. Psychol.* 12, 1–12. doi: 10.3389/fpsyg.2021.808525
- Kress, G., and van Leeuwen, T. (2006). *Reading Images: The Grammar of Visual Design*. London: Routledge. doi: 10.4324/9780203619728
- Leung, D., Law, R., van Hoof, H., and Buhalis, D. (2013). Social media in tourism and hospitality: a literature review. *J. Travel Tour. Market.* 30, 3–22. doi: 10.1080/10548408.2013.750919
- Maci, S. M. (2020). *English Tourism Discourse: Insights into the Professional, Promotional and Digital Language of Tourism*. Milan: Hoepli Editore.
- Mai, L., Le, H., Niu, Y., and Liu, F. (2011). “Rule of thirds detection from photograph,” in *2011 IEEE International Symposium on Multimedia* (Dana Point, CA: IEEE), 91–96. doi: 10.1109/ISM.2011.23
- Manovich, L. (2017). *Instagram and Contemporary Image*. Available online at: [www.manovich.net](http://www.manovich.net) (accessed December 10, 2023).
- Mattei, E. (2023a). Theory and method for the statistical investigation of multimodal promotional practices in the digital era: a data-driven approach based on systemic functional linguistics and social semiotics. *IDEAH* 3, 1–32. doi: 10.21428/f1f23564.7921b725
- Mattei, E. (2023b). *Multimodal Corpus Analysis of Digital Tourism Narratives: A Data-Driven Approach Based on Systemic Functional Linguistics and Social Semiotics* (Dissertation). University of Verona. Available online at: <https://iris.univr.it/handle/11562/1098826> (accessed February 5, 2024).
- Mattei, E. (2023c). “Investigating multisemiotic persuasive practices by integrating computational methods and complementary theoretical frameworks,” in *A Data-driven Approach to Digital Tourism Discourse Based on Systemic Functional Linguistics and Empirical Multimodality. Digital Humanities 2023. Collaboration as Opportunity (DH2023)*, eds W. Scholger, G. Vogeler, T. Tasovac, A. Baillot, and P. Helling (Austria: Centre for Information Modeling), 1–2. doi: 10.5281/zenodo.8210654
- Mattei, E. (forthcoming). “Integrating computational and statistical methods into the humanities: investigating multimodal tourism discourse with empirical social semiotics and SRI tagging software” in *Language, Data Science and Digital Humanities*, eds M. Laitinen, and J. Tyrkkö (Finland: University of Eastern Finland; Sweden: Linnaeus University).
- Messaris, P. (1997). *Visual Persuasion: The Role of Images in Advertising*. London: Sage.
- Mitchell, W. J. T. (1994). *Picture Theory: Essays on Verbal and Visual Representation*. Chicago, IL: University of Chicago Press.
- Pibiri, G. E., and Mattei, E. (2020). *Statistically Reliable Image Tagging*. Available online at: <http://xor.isti.cnr.it:8000/login.html> (accessed January 30, 2024).
- Shuqair, S., and Cragg, P. (2017). The immediate impact of instagram posts on changing the viewers’ perceptions towards travel destinations. *Asia Pac. J. Adv. Bus. Soc. Stud.* 3, 1–12. doi: 10.25275/apjabssv3i2bus1
- Smith, S. P. (2021). Landscapes for “likes”: capitalizing on travel with Instagram. *Soc. Semiot.* 31, 604–624. doi: 10.1080/10350330.2019.1664579
- Sormaz, A., and Ruoss, E. (2020). “Social media to balance tourism flow in natural heritage destinations,” in *Proceedings of the Heritage, Tourism and Hospitality International Conference “Living Heritage and Sustainable Tourism”*, eds L. Cantoni, S. De Ascaniis, and K. Elgin-Nijhuis (Lugano: Università della Svizzera Italiana), 15–28.
- Stöckl, H., Caple, H., and Pflaeging, J. (2020). *Shifts Towards Image-centricity in Contemporary Multimodal Practices*. London: Routledge.
- Tomlinson, T. (2007). *The Culture of Speed: The Coming of Immediacy*. London: Sage.
- Urry, J., and Larsen, J. (2011). *The Tourist Gaze 3.0*. London: Sage Publications.





## OPEN ACCESS

## EDITED BY

Janina Wildfeuer,  
University of Groningen, Netherlands

## REVIEWED BY

Tuomo Hiippala,  
University of Helsinki, Finland  
Hartmut Stöckl,  
University of Salzburg, Austria

## \*CORRESPONDENCE

Jacopo Castaldi  
✉ jacopo.castaldi@canterbury.ac.uk

RECEIVED 10 November 2023

ACCEPTED 06 March 2024

PUBLISHED 19 March 2024

## CITATION

Castaldi J (2024) Refining concepts for empirical multimodal research: defining *semiotic modes* and *semiotic resources*. *Front. Commun.* 9:1336325. doi: 10.3389/fcomm.2024.1336325

## COPYRIGHT

© 2024 Castaldi. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Refining concepts for empirical multimodal research: defining *semiotic modes* and *semiotic resources*

Jacopo Castaldi\*

Centre for Language and Linguistics, Canterbury Christ Church University, Canterbury, United Kingdom

The issue of defining key concepts in multimodal research is at the same time ongoing and of pivotal importance. Building on John Bateman's categorisation of modes, and paying special attention to the concept of *materiality* within the discussion, the paper provides a clear differentiation between *semiotic modes* and *semiotic resources* and discusses the relationship between the two. These will be defined by also looking at how they differ from another key concept in multimodal research, i.e., *media*, and examples will be provided to illustrate how the newly defined concepts can guide empirical investigations of multimodal texts and their reception. The paper aims to continue the discussions around these key concepts amongst multimodal scholars, so that agreement in the field can eventually be reached.

## KEYWORDS

**semiotic modes, semiotic resources, media, materiality, empirical multimodality research**

## 1 Introduction

The issue of defining key concepts in multimodal research is at the same time ongoing and of pivotal importance. This paper aims to contribute to ongoing discussions by offering a clear differentiation between *semiotic modes* and *semiotic resources* as well as a discussion of the relationship between the two. Although the concept of semiotic mode is of key importance to multimodal research, a review of the literature in the field shows, at best, contrasting definitions and, at worst, the suggestion that a clear understanding of what modes are may be of no use at all. The last stance is the one taken by Machin (2013) who asserts that, since it has proved very difficult to ascertain what constitutes modes, "MCDS (Multimodal Critical Discourse Studies) may turn out to have less use with the issue of what modes are in themselves as with how different kinds of semiotic resources can play a part in realising discourses since they are good at doing different things" (p. 349). Notwithstanding the importance of the last part of his assertion, I believe it may be equally difficult to establish what different things modes are good at if we do not first establish what they are.

A similar point can be made for the concept of *semiotic resources*, which is sometimes seen as an overall umbrella term for anything that can be used for meaning-making, and whose nature and composition is often vaguely defined. Indeed, as Bateman (2021a, p. 56) states:

Considerable theoretical uncertainty therefore remains concerning just how potentially “overlapping” semiotic systems might best be approached, both theoretically and practically during analysis. This is not helped by the fact that the notion of “semiotic resource” is also intrinsically vague — anything that may serve a semiotic purpose may be a resource: van Leeuwen even writes, for example, of “genre” being a semiotic resource (Van Leeuwen, 2005: 128). This does not provide support for empirical analysis.

The starting point for clarifying the ontological status of modes and resources, as well as the relationship between the two, will be the definitions of semiotic modes provided by Bateman (2011, 2016) and Bateman and Schmidt (2012). After revisiting some of the literature definitions around the concepts of *modes* and *resources*, a proposal is put forward to differentiate between the two. The differentiation is based on the (relatively) stable material properties of semiotic modes and on the ability of semiotic resources to be deployed and articulated through different materialities. Semiotic resources are then categorised in this present paper by drawing on the Systemic Functional Linguistic (henceforth SFL) *ideational*, *interpersonal* and *textual* metafunctions (e.g., Halliday, 1978), and four dimensions are discussed: discursive resources, pragmatic resources, stylistic resources and textual resources. Finally, a proposal is put forward to place semiotic resources at an intermediate stratum between paradigmatic and syntagmatic axes of organisation, and discourse semantics within the composition of a mode. The reason for this, the paper argues, is that this intermediate stratum will help explain *how* the semiotic codes, by which I refer to the first two strata in Bateman’s (2011) model, take the specific configurations that allow to activate certain interpretations (and not others) at the final stratum of discourse semantics.

A discussion about semiotic modes and semiotic resources, however, cannot do without addressing another key concept in multimodal research, namely media. By looking at all three concepts, i.e., modes, resources and media, a central role is attributed to the material dimension of signification and interpretation. On the one hand, the paper argues that materiality is a key constitutive component of modes and media, both of which rely on relatively stable and historically developed material substances. On the other hand, it is argued that materiality, despite gaining importance once semiotic resources are deployed in actual texts, does not represent a constitutive element of semiotic resources, which are instead defined as abstract metafunctional constructs that can be realised through different materialities and/or semiotic codes.

The paper begins by discussing and defining semiotic modes and semiotic resources as well as by clarifying the relationship between the two. Since, however, “we do not find ‘free-floating’ instances of semiotic modes,” media will also be discussed, as they “group semiotic modes dynamically into socioculturally and historically situated configurations” (Bateman, 2017, p. 168). The role of materiality for multimodal research will then be discussed in order to establish two parallel lines of empirical enquiry. The first is how the materiality of the signs and sign systems affects their deployment in communication; the second is how socio-cultural conventions, as well as technological advancement, shape and alter the range of material configurations that can be deployed through specific modes, media and genres. These will be further explored in the following section, which outlines implications for empirical multimodal research and offers pointers for potential research endeavours that can focus both on text production

and text reception at three different levels of analysis (cf. Bateman, 2021b, pp. 3–4): (i) investigating which semiotic resources and metafunctions individual modes can actualise; (ii) investigating the relationship between different modes actualising the same semiotic resources and metafunctions; (iii) investigating the contribution of individual modes to perform the three metafunctions of a communicative event.

## 2 Semiotic modes

The most problematic issue with defining and categorising modes seems to be the difficulty to establish clear boundaries between them (Machin, 2013, p. 349). Within many approaches to multimodality, however, modes have been generally equated to systems of signs, e.g., speech, writing, gestures, sounds, etc. (Kress and van Leeuwen, 2001, p. 6; p. 9; Forceville, 2009, p. 23; Page, 2009, p. 6; Kress, 2010, p. 79; Stöckl, 2014). Kress (2010, p. 87) identifies both a *social* and a *formal* dimension of modes, with the former relating to specific communities and their contingent “social-representational needs,” and the latter aligning with Halliday’s (1978) three metafunctions (i.e., ideational, interpersonal and textual). The formal dimension in Kress’ formulation (built on Kress and van Leeuwen, 1996), however, has been criticised on the grounds that not all modes seem to be able to fulfil all the three metafunctions (Van Leeuwen, 1999, pp. 190–191; Bateman, 2021b, p. 4) and that not all modes can and should be treated in the way language has within the SFL tradition (Machin, 2016, p. 327).

One notable exception to the equation of modes with sign systems is O’Halloran (2005, pp. 20–21) who maintains that modes are related to the sensory channels of communication, while defining the systems of signs as semiotic resources. The latter view, however, has been criticised on the grounds that “modes cut across sensory channels, so the nature of a sign is not sufficiently characterised by looking at its path of perception” (Stöckl, 2014, p. 11). A similar point is made by Bateman (2021a, pp. 49–50), who also highlights how the “conflation of the material and the semiotic, mak[es] analysis and demarcation of data unnecessarily complex.” In agreement with Stöckl and Bateman, the five senses of seeing, hearing, touching, smelling and tasting will be understood in this paper to refer to *sensory channels* and not modes.

Notwithstanding the need to avoid conflating the material and the semiotic, Bateman (2011) and Bateman and Schmidt (2012) claim that the materiality of the medium as well as that of the systems of signs, need to be taken into consideration for a full account of what modes are and therefore do not provide a list of modes but, rather, a breakdown of the layers of a “three-stratal organisation” that comprise modes, namely (i) a material substrate; (ii) paradigmatic and syntagmatic axes of organisation (e.g., a lexicon and a grammar in the case of language); (iii) a discourse semantics through which the ‘semiotic code’, defined as the combination of (i) and (ii) above, becomes interpretable, and hence a “fully fledged semiotic mode” (Bateman, 2011, pp. 20–22).<sup>1</sup> Kress and van Leeuwen (2001) also stress the importance of the discursive dimension of modes by stating that

1 From this point onwards, following Bateman’s definition, a “semiotic code” is meant to refer to a *context-potential* sign system, whereas a “semiotic mode” is meant to refer to a *context-actual* sign system.

these “allow the simultaneous realisation of discourses and types of (inter) action” (p. 21).

The material substrate in the first stratum allows the signs and the sign systems to be perceivable and, at the same time defines them. As for the second stratum, i.e., “paradigmatic and syntagmatic axes of organisation”, Bateman and Wildfeuer (2014, p. 186) maintain that “this ‘mid-level’, or mediating stratum generally operates compositionally and can be characterized independently of context.” The final stratum, discourse semantics, provides the connecting ‘tissue’ between the “somehow ‘interpretable’ in context” (Bateman, 2011, p. 21) “semiotic code” and its situated communicative enactment by “provid[ing] the *pragmatic interpretative mechanisms* necessary for relating the forms a semiotic mode distinguishes to their contexts of use and for demarcating the *intended range* of interpretation of those forms” (*ibid*, p. 181, *emphasis in original*). It can be argued, however, that within this definition it remains unclear just *how* the paradigmatic and syntagmatic structures in the first two strata take the specific forms that constrain the interpretative options at the level of discourse semantics. The latter relies on elements that are already beyond the materiality of the semiotic code, e.g., genre recognition, recognition of metafunctions, cultural understanding and so on, and that, moreover, are very often performed by the *combination* of modes in complex multimodal artefacts (Bateman, 2021b, p. 4).

Two aspects of Bateman and colleagues’ work, however, already include a potential solution to the problem outlined above, and this is the use of the word “resources” at different stages in the development of the model to refer to the second stratum of their classification. Bateman (2011, p. 20) uses the concept of *semiotic resources* to describe “semiotically charged organisations of material that can be employed for sign-construction,” which in his theorisation equates to the second tier of the “three-strata organisation” of semiotic modes. More recently, Hiippala and Bateman (2021, pp. 407–408) refer to the second stratum as *expressive resources*, which are “assumed to be subject to a paradigmatic organization that allows making selections among them and combining them into larger syntagmatic organizations” and examples of which, within the context of a diagrammatic mode, are written language and line drawings. My proposal is that the construction and structuring of syntagmatic and paradigmatic resources within the actual deployment of a mode are not only constrained by the material qualities of the stratum below but also by abstract, *potential* metafunctional constructs that guide paradigmatic and syntagmatic choices and take *specific forms* at the discourse semantics stratum to activate certain interpretations and not others. An example is necessary at this stage to illustrate this line of thinking.

If we take the written language mode as deployed in an academic article, there will be a number of metafunctional considerations that will influence both the paradigmatic and syntagmatic levels in the second stratum of the model. Stylistic considerations, for example, will guide the choice of lexicon amongst the paradigmatic options by selecting more formal lexical options; likewise, at the syntagmatic level, certain syntactical structures will be preferred by selecting, in the case of English academic writing, subordinate clauses and passive voices. These choices are at the same time not a defining feature of the paradigmatic and syntagmatic structures of the written language code nor dictated by the internal structure of the code, but by the external socio-cultural expectations connected with the deployment of the mode for a specific communicative purpose. The stylistic

metafunctional constructs considered at the paradigmatic and syntagmatic level, once selected and realised, will enable a restricted number of interpretations at the final level of discourse semantics.

However, if we accept the necessity to include this set of metafunctional constructs somewhere between the material substrate and the discourse semantics stratum, the question arises of where they should be placed in the model. One option is that the ‘expressive resources’ in the second stratum are expanded to include the metafunctional constructs alongside paradigmatic and syntagmatic options; another is to posit a further stratum that sits between the second and third strata of the model, thus creating a four-stratal model. Before being able to answer this question, however, the exact nature of these ‘resources’ needs to be established as well as their relationship with the strata as they currently appear in Bateman’s model. The following section will argue that the concept of *semiotic resources*, once clearly defined, can provide the answer to the question above.

### 3 Semiotic resources

As Van Leeuwen (2005, p. 3) states, the idea of semiotic resources is taken from Halliday’s SFL, in which grammar is described as a “resource for making meanings” (Halliday, 1978, p. 192). Van Leeuwen then goes on to give a detailed description of what this means:

In social semiotics resources are signifiers, observable actions and objects that have been drawn into the domain of social communication and that have a *theoretical* semiotic potential constituted by all their past uses and all their potential uses and an *actual* semiotic potential constituted by those past uses that are known to and considered relevant by the users of the resource, and by such potential uses as might be uncovered by the users on the basis of their specific needs and interests. Such uses take place in a social context, and this context may either have rules or best practices that regulate how specific semiotic resources can be used, or leave the users relatively free in their use of the resource (Van Leeuwen, 2005, p. 4; *emphasis in original*).

The way semiotic resources are defined in the quotation above means they encompass pretty much anything that can be used for meaning-making, provided that they are one of possible options from which users can choose and that they can be used following a more or less strict set of rules. It is for this reason that Kress and van Leeuwen (2001, pp. 21–22) suggest that not only modes but also media are examples of semiotic resources, once the “principles of semiosis [of media] begin to be conceived of in more abstract ways (as ‘grammars’ of some kind).” Despite the fuzziness of the boundaries of semiotic resources in van Leeuwen’s definition, and hence the difficulty to apply the concept in empirical investigations, it is important to note the point that the intended context of use will influence the choice of resources to be deployed, a point to which I will return later in this section.

Unlike Kress and van Leeuwen, O’Halloran (2005, p. 20) does not include media amongst semiotic resources and lists “speech, music and diegetic sound” (in effect what almost everyone else defined as modes) amongst examples of semiotic resources.

Bateman (2011, p. 20) defines semiotic resources as “semiotically charged organisations of material that can be employed for sign-construction,” which, as we have seen in the previous section, equates to the second tier of their three-stratal organisation of semiotic modes. Machin and Mayr (2012), finally, do not define semiotic resources as such, but talk about lexical and visual repertoires, which are the two dealt with in their book, in lieu of semiotic resources. There is, therefore, either considerable overlap between modes and semiotic resources to the point that one of the terms becomes redundant, or a lack of clear boundaries, which conflates very different concepts under the same broad umbrella of *meaning-making*.

As Bateman (2021a, p. 55) notes, however, the conflation of semiotic modes with a broader notion of semiotic resources “results in ‘semiotic mode’ saying little more than is already covered by the term ‘semiotic resource.’” I would argue, therefore, that the effort to define semiotic modes has to be coupled with the effort to provide a clear definition and classification of semiotic resources. Providing clear-cut, discrete categories and constructs within the broad (and vague) umbrella of ‘resources’ allows researchers to focus empirical investigations, as I will show in section 6. It has to be stressed at this point that the categorisation of semiotic resources that follows is embedded *within* the composition of a semiotic mode, while, following van Leeuwen’s definition of semiotic resources quoted earlier, being also affected by the intended context of use in which the modes will be deployed.

To begin our categorisation of *semiotic resources* as part of a mode, I would argue that these can be defined as abstract, potential metafunctional constructs that can be realised through different materialities and/or semiotic codes. Bateman (2021a, p. 49) theorises a similar “‘abstract’ or ‘generalised’ materiality” when he discusses the concept of *canvas*. A canvas is defined as the materiality of a semiotic mode “*when viewed with respect to the specific forms of traces required by that semiotic mode*” (Bateman, 2021a, p. 46, *emphasis in original*). A parallel can be drawn with semiotic resources as they, too, albeit already existing as abstract constructs, will take different materialities and leave different traces, depending on the semiotic codes deployed to actualise them. It has to be noted at this point that the constructs I propose should be regarded as ‘code-dependent contributions of resources’ to communicative metafunctions and not as the overall final actualisation. For example, different modes will contribute different aspects to the genre of *lectures*, but their contribution will be guided by the semiotic resources component *within* the mode composition and, in turn, limit the possible genre recognition options at the level of discourse semantics (see further below). Now that an initial definition of semiotic resources has been offered, it is possible to address the question raised at the end of the previous section, that is whether semiotic resources should be placed with paradigmatic and syntagmatic structures in the second stratum of Bateman’s model, or whether a further stratum should be added to the model, thus making it a four-stratal one. Following an answer to this issue, I will then provide an initial, tentative categorisation of semiotic resources as belonging to four dimensions, *discursive*, *pragmatic*, *stylistic* and *textual*, based on SFL metafunctions.

I propose that the second option, that is an additional independent stratum that sits between paradigmatic and syntagmatic axes of organisation and the final stratum of discourse semantics, should be adopted for the following reasons. Firstly,

semiotic resources, as I defined them, represent abstract, a-material options that are not dependent on the materiality of the first two strata, but that can take different forms in relation to the materiality and related affordances of the first two strata. *They can therefore be applied, in an abstract fashion, to any semiotic codes*. However, some resources and related metafunctions may not be available at all to some codes: Van Leeuwen (1999, cf. Bateman, 2021b, p. 4), for example, problematises the idea of modes being able to fulfil all metafunctions:

Looking back I would now say that different semiotic modes have different *metafunctional configurations*, and that these metafunctional configurations are neither universal, nor a function of the intrinsic nature of the medium, but cultural, a result of the uses to which the semiotic modes have been put and the values that have been attached to them. Visual communication, for instance, *does* have its interpersonal resources, but they can only be realized on the back of ideation, so to speak. If you want to say ‘Hey you, come here’ by means of an image, you have to do it by representing someone who makes a ‘Hey you, come here’ gesture. You cannot do it directly. With sound it is the other way around. Sound *does* have its ideational resources, but they have to be realized on the back of interpersonal resources (Van Leeuwen, 1999, pp. 190–191, *emphasis in original*).

Furthermore, combinations of these resources can result in other communicative constructs, such as rhetorical strategies, which Bateman (2014, p. 250) defines as “some binding of, on the one hand, communicative ‘goals’ [...] and, on the other hand, selected realisation strategies ranging over any of the semiotic modes that can be mobilised in an artefact.” There is therefore an ontological difference between the materiality of the paradigmatic and syntagmatic structures on the one hand, and the a-materiality of the semiotic resources, together with their potential combination into communicative strategies, on the other, which would be best reflected by placing them in a separate stratum.

Second, an intermediate stratum between the bottom two (material substrate+paradigmatic and syntagmatic axes of organisation) and the top one (discourse semantics) is necessary to negotiate *at an abstract level* between two set of sign-making forces: the material affordances of the code on the one side, and the socio-cultural expectations surrounding the mode deployment through specific media and genres on the other. This abstract negotiation is then actualised in specific, material, interpretable discourse semantics. For example, if we posit that the stratum of discourse semantics can include pragmatic competence such as genre recognition (Bateman, 2017), then it necessarily follows that the mode must have had access to relevant pragmatic options (or any other metafunctional constructs) at a lower stratum so that the interpretability of the mode at the level of discourse semantics can be ‘activated’ and hypotheses can be generated which involve ascribing them to particular semiotic resources (Hiippala and Bateman, 2022, p. 16). This process of hypothesis generation and identification of semiotic resources must include both a stage where the intended ones are selected from the options at the lower stratum (i.e., the proposed additional stratum of semiotic resources), and a stage where an intended audience is able to generate abductive hypotheses based on the material traces of those selections, as crystallised in the final stratum of discourse semantics.



The choice of which semiotic resources to draw on can therefore happen through both bottom-up (text producer-driven) and top-down (context-driven) factors, or, indeed and perhaps more likely, a combination of both. Bottom-up, the range of resources that can be accessed will depend on the material codes, and related affordances, at the disposal of the text producer; within the accessible range of semiotic resources for the codes selected, individual, stylistic preferences of the sign-maker may also influence the selection of specific resources. Top-down, the choice of resources will be influenced by the contingent socio-cultural expectations related to the genres and media through which communication occurs. These socio-cultural expectations will also guide the correct interpretation of the multimodal artefact based on the specific discourse semantics resulting from the contribution of the co-occurring modes.

Let us provide some examples of both processes at work, beginning with the bottom-up scenario and considering the institutional practice of *giving a lecture*. The text-producer will have, depending on the technology available in the classroom, a number of modes they can choose from, including spoken language, written language as deployed in hand-outs and digital presentation material, images, diagrams and other visuals as deployed in hand-outs and digital presentation material, to mention some of the most commonly used, say, in British Higher Education. We can focus on one of these modes, the *spoken language*, and on one of the pragmatic semiotic resources necessary to fulfil the purpose of 'enabling teaching and learning', that of *text types*.<sup>2</sup> The text producer will choose to activate those text types that are deemed to be functional to that purpose, i.e., informative and descriptive. These will then require specific syntagmatic and paradigmatic configurations available at the lower stratum and, together, form the material basis of a specific discourse semantics which will allow, on the one hand, the text producer to construct the multimodal communicative act by harmonising the various contributing modes, and, on the other, will allow the students to recognise and interpret (not necessarily at a conscious level) the meaningful contributions of the individual modes for the purpose of teaching and learning. If, alternatively, we take the related but different practice of *academic conference presentation*, the text producer, through the spoken language mode (and indeed other co-occurring modes) will necessarily make use of other text types in addition to the informative and descriptive, i.e., the argumentative, that are necessary to fulfil the purpose of convincing an audience that the propositions advanced in the presentation are to be accepted as valid. One could argue that argumentative text types will also be necessary in a lecture and, to a certain extent, this is probably the case. However, given the audience (students) and their expectations of the practice (learning, developing intellectual skills), as well as the relative position of power of the 'expert' lecturer vis-à-vis the 'novice' student, the choice of text types will be skewed towards the informative and descriptive rather than towards the argumentative. The opposite would apply in the practice of conference presentation, where the audience (peers) and their expectations (being presented with something original and

scientifically tenable) will skew the text type proportions towards the argumentative one.

The final observations regarding audiences and their expectations take us already in the realm of top-down processes, that is the context-driven ones. Beside considerations around audiences and their expectations, the limitations imposed by institutionalised practices will also influence the choice of semiotic resources the codes are allowed to activate. This is also in line with the point highlighted at the beginning of this section in van Leeuwen's broad definition of semiotic resources, that is that the intended context of use will influence the choice of semiotic resources (cf. also Bezemer, 2023, p. 16). Since the discourse semantics stratum involves situated and contextual discourse interpretations (Bateman, 2011, p. 22), the options of semiotic resources to deploy in such situated and contextual discourses will also be influenced by the cultural and institutional limitations posed by specific contexts. We can explicate the top-down process with the institutional practices introduced above but this time we will work through the model backwards (i.e., from the higher strata to the lower ones). The *lecture* social practice, which can also be seen as a genre, is not the only available option to fulfil the purpose of 'enabling learning and teaching', *seminars* and *workshops* being other notable alternatives. However, the institutional practice may impose restrictions on which genre to be used at a specific time and place, as may do the technology available. In turn, the (imposed) choice of a lecture over a seminar or workshop will require a specific discourse semantics that needs to allow for specific interpretations, for example concerning the role of the participants, their relative power in the proceedings, interactional turns and so on. The desired discourse semantics will then draw on the best suited abstract semiotic resources available at the lower stratum and these will take specific forms depending on the semiotic codes that can be utilised depending on the canvas(es) available. Whether or not the choice of the semiotic resource is driven by bottom-up (i.e., text producer) or top-down (i.e., contextual) factors is irrelevant with regard to the status of semiotic resources, that is a set of abstract, a-material options available to perform metafunctions.

A similar line of reasoning can be followed for the other dimensions of the semiotic resources as I will categorise them. However, the examples provided should suffice to explicate the ontological status of semiotic resources as a set of abstract, a-material options available in relation to specific codes as well as the need to be placed at an intermediate stratum between paradigmatic and syntagmatic properties, and discourse semantics. Put differently, what I maintain is that the stratum of semiotic resources provides the metafunctional coordinates that dictate a certain organisation of the axes. So, although they are a-material, they guide the structuring of the syntagmatic and paradigmatic axes so that only certain *actual* interpretations can be made available at the level of discourse semantics. Without these set of coordinates, i.e., the semiotic resources as defined here, we are missing the link that allows to move from all the possible paradigmatic and syntagmatic structural options available to a code to the actual ones that lead to a specific discourse semantics. At the level of discourse semantics, the paradigmatic and syntagmatic structures have already taken *specific material configurations*, thus achieving an ontological status that is the sum of strata (ii) (paradigmatic and syntagmatic axes of organisation) and (iii) (semiotic resources). These new material, metafunctionally-loaded configurations at the level of discourse semantics need to be necessarily

<sup>2</sup> Text types, following the German school of text linguistics applied to translation (Nord, 1991, p. 18), refer to narration, report, description, exposition and argumentation.



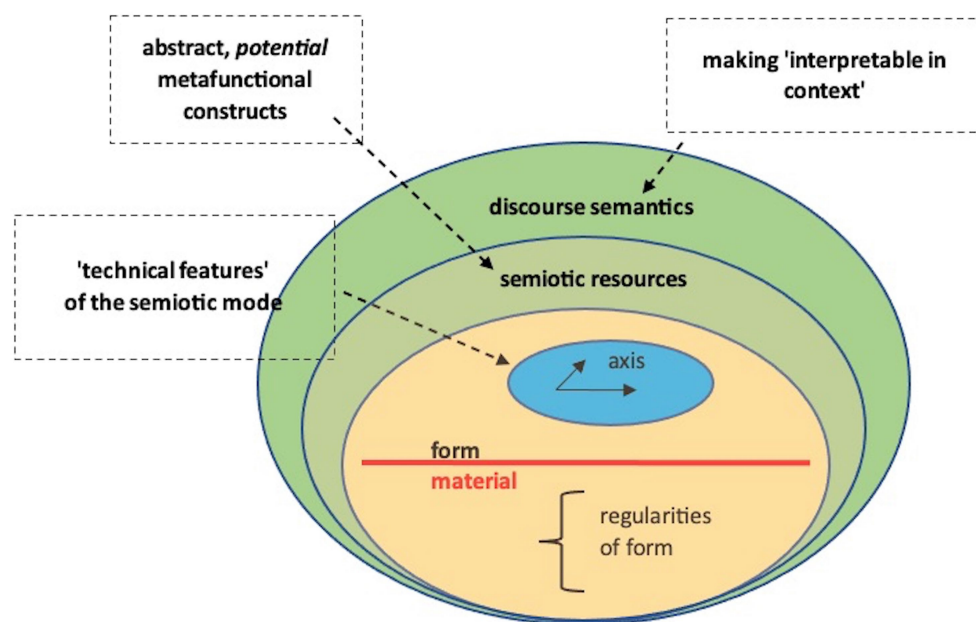


FIGURE 1  
Visual schematisation of the expanded theoretical model of mode (after Bateman, 2016).

fixed (and no longer potential and a-material). Without a fixed materiality it would not be possible for modes to co-occur based “on the affordances of the materialities being combined” within a specific medium (Bateman, 2016, p. 56). This material ontology of the individual discourse semantics of modes is also highlighted by Bateman (2011, p. 27, *my emphasis*):

Providing a formalised account of the kinds of semantics that applies for each semiotic mode, together with a close mapping between properties of the *articulated material* and those semantics, is the first step towards a well-founded account of the semantics of modes both individually and in combination.

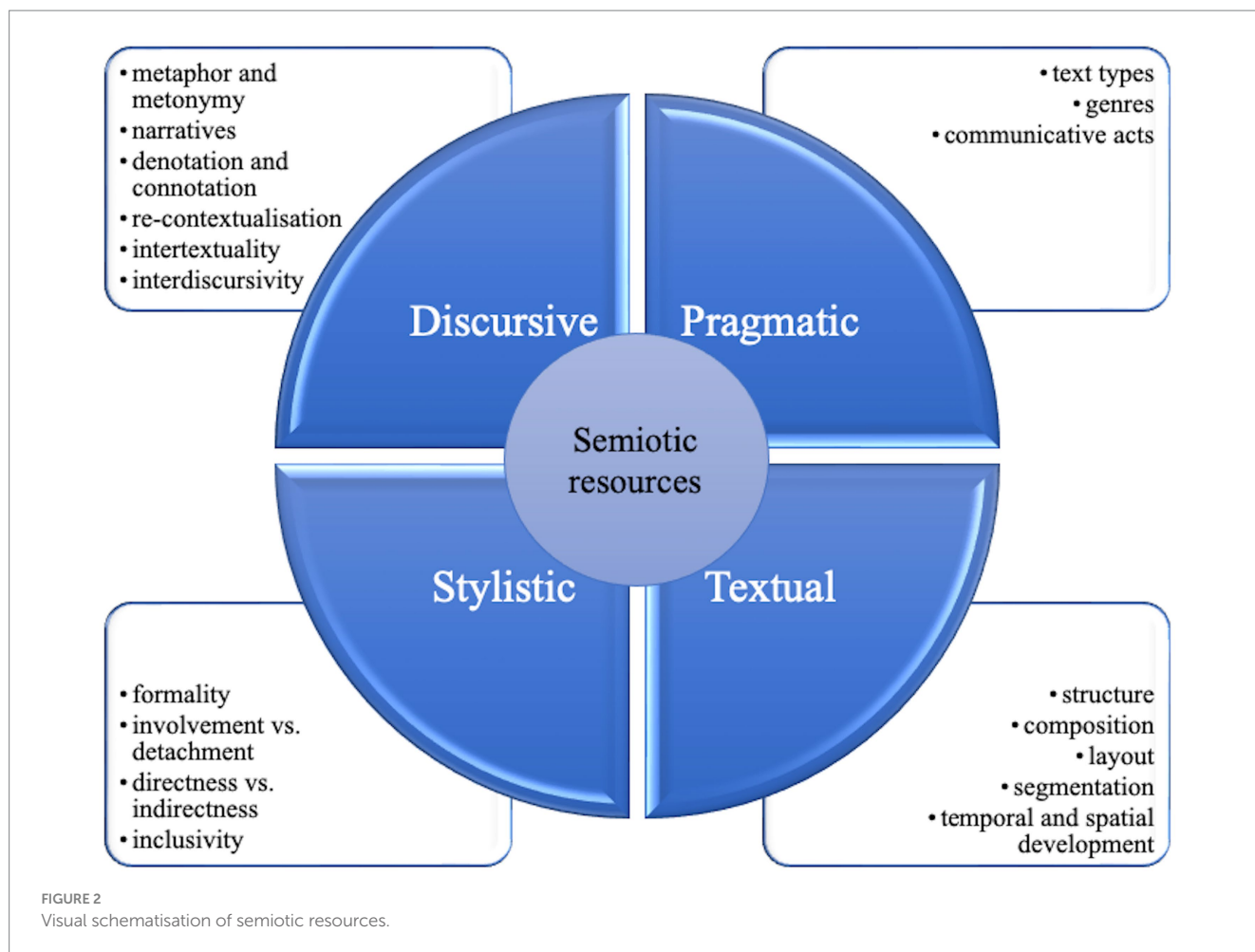
It has to be stressed, however, that the intermediate stratum I propose does not change the overall structure theorised by Bateman (2016), which sees the communicative event formed by modes interacting within specific media and being finally attributable to a genre that is recognisable by the participants in the communicative event. Based on the discussion so far, a four-stratal organisation of semiotic modes can be offered, which builds on the models discussed so far and comprises: (i) a material substrate; (ii) paradigmatic and syntagmatic axes of organisation; (iii) semiotic resources; (iv) discourse semantics. Figure 1 provides a schematic visualisation, based on Bateman (2016), of the expanded theoretical model of mode.

Now that the ontological status of semiotic resources and their relationship to modes have been established, we can move on to provide a finer categorisation of semiotic resources. To this purpose, I propose to arrange them into four macro areas: discursive, pragmatic, stylistic and textual. Discursive resources allow conceptualisation: they primarily attend to the *content* of communicative events and can be roughly equated with the SFL *ideational* metafunction. Examples of discursive resources include metaphor and metonymy, narratives, denotation and connotation, re-contextualisation, intertextuality and interdiscursivity. Pragmatic resources allow purpose: they primarily attend to the *function* of communicative events. Examples are text

types (narration, report, description, exposition and argumentation), genres (travel documentaries, sci-fi films, etc.) and communicative acts (e.g., invitation, offer, command, request, etc.).<sup>3</sup> Stylistic resources allow *agency*: they primarily attend to *identities* in communicative events. Examples are formality, involvement vs. detachment, directness vs. indirectness and inclusivity. Pragmatic and stylistic resources can be roughly equated with the SFL *interpersonal* metafunction. Finally, textual resources allow *organisation*: they primarily attend to the *structure* of communicative events and can be roughly equated with the SFL *textual* metafunction. Examples of textual resources include structure, composition, layout, segmentation, temporal and spatial development. Equating SFL metafunctions to the semiotic resources rather than to the semiotic modes gives the theoretical advantage to be able to account for those semiotic modes that do not present all three metafunctions (Van Leeuwen, 1999, pp. 190–191; Bateman, 2021b, p. 3), since these properties are now part of the semiotic resources. Figure 2 is a schematic representation of the *semiotic resources* as defined above, but the lists within each area should not be taken as exhaustive.

To summarise, this section aimed to provide clarifications and definitions on three fronts. First, it discussed the nature of semiotic resources and defined them as *abstract potential metafunctional constructs*. Second, it argued that a further stratum should be added to the theoretical definition of mode provided by Bateman and colleagues to accommodate the newly defined concept of semiotic resources, and discussed the relationship between the new stratum and those below and above it. Finally, it provided an initial, tentative categorisation of semiotic resources by grouping them under the categories of discursive, pragmatic, stylistic and textual and by

<sup>3</sup> I am using *communicative acts* in place of the most commonly used *speech acts* to extend this pragmatic concept to non-linguistic modes (see also Bucher's (2017, p. 110 ff.) definition of multimodality as *communicative action*).



equating them to the SFL ideational, interpersonal and textual metafunctions. The next section will look at the concept of media and discuss the relationship between semiotic modes, semiotic resources and media.

## 4 Media

A first, mostly agreed upon, distinction is made between *modes* and *media* and, accordingly, between *multimodality* and *multimediality*, with the former referring to the simultaneous deployment of different modes and the latter to the simultaneous deployment of different media. Kress and van Leeuwen (2001) refer to media as “the material resources used in the *production* of semiotic products and events, including both the tools and the materials used (e.g., the musical instrument and air; the chisel and the block of wood)” (p. 22, *my emphasis*) and connect media to the sensory system (*ibid*, p. 67). O’Halloran (2005), on the other hand, focuses on the *distribution* and *reception* of media, by defining them as the “material resources of the channel” and presenting, as examples, platforms such as the radio and websites (p. 20). Elleström (2010) offers a very sophisticated view of media and rejects the idea of modes being “[e]ntities such as ‘text’, ‘music’, ‘gesture’ or ‘image’” (p. 16). He sees media as the starting point and maintains that, in order to fully appreciate and analyse how media work, one needs to consider four different *modalities* that are all necessary conditions for any medium to exist: a

*material modality*, a *sensorial modality*, a *spatiotemporal modality* and a *semiotic modality*. These “are to be found on a scale ranging from the tangible to the perceptual and the conceptual” (Elleström, 2010, p. 15) and, although not chronologically or hierarchically ordered, they can be approached in that order as each modality depends on the existence of the previous one to be accessed (*ibid*, p. 17). Materiality is therefore a defining feature of media and the latter, following Elleström (2010), can be defined as the material channels, be these animate or inanimate, through which communicative events are produced, distributed and received.

Of particular interest to our discussion is the *semiotic modality*. This modality attends to meaning, with the latter to “be understood as the product of a perceiving and conceiving subject situated in social circumstances” (Elleström, 2010, p. 21). The semiotic modality is what allows people to interpret signs through two different ways of thinking: an abstract one directed by *propositional* representations “created by conventional, symbolic sign functions,” that is signs that have no resemblance or association with the object they refer to (e.g., words or a red light to imply “stop”);<sup>4</sup> a direct one directed by *pictorial* representations “created by indexical and iconic sign functions” (*ibid*,

<sup>4</sup> Although some words can be described as “iconic symbols” (e.g., onomatopoeic words) and “indexical symbols” (e.g., deictic words) (Chandler, 2017, p. 56).

p. 22), that is signs that have an association with the object they refer to (an *index*, e.g., smoke signalling a fire) or that refers directly to the object (an *icon*, e.g., a photo or an emoji). Using terminology from Peirce's semiotics, Elleström (2010, p. 22) therefore suggests "that convention (symbolic signs), resemblance (iconic signs) and contiguity (indexical signs) should be seen as the three main modes of the semiotic modality".

Elleström's discussion is centred around the focal concept of *medium* and the term *modalities* can create confusion in, for example, a social semiotic approach to multimodality where modality is used to refer to the degree of epistemic value of the signs (e.g., Van Leeuwen, 1999, p. 170).<sup>5</sup> Despite the terminological confusion, I believe that his unpacking of what makes media what they are is compelling, as it touches on all the elements (materiality, senses, cognition and semiosis) that need to be considered in a multimodal approach to communication, particularly if the interaction of an audience with the media is also analysed. There are, however, some issues with Elleström's all-encompassing definition of media, particularly when it comes to understanding the relationship between media and modes. Bateman (2017, p. 168) maintains that the primary role of media is to "provide the immediate context in which semiotic modes can be used." He therefore argues that the relationship between media and modes is not one of interdependence and highlights how.

On the one hand, semiotic modes are always more 'local' organisations that take responsibility for the deployment of specific material regularities. They are definitionally independent of media. On the other hand, media are broader 'second-order' phenomena constituted by socioculturally specific bundlings of semiotic modes and, as a consequence, may not be directly perceptible in their own right (p. 172).

With reference to the role of *semiotic resources* as defined in section 3, a similar line of thought can be followed, first and foremost because semiotic resources are now defined as a constitutive component of semiotic modes. The representational force afforded by semiotic resources in the process of deploying modes in situated communicative contexts is also fully realised only when ascribed to a specific medium and, as per Bateman (2016), to a specific genre. As we have argued before, it is often the combination of modes with their individual metafunctional possibilities and limitations that, through the higher-order levels of media and genre, allows a communicative event to be able to realise *all* three metafunctions and therefore acquire full communicative effectiveness.

Now the hierarchical relationship between modes (which include semiotic resources) and media has been established we can turn to discuss in more detail the role of materiality in the analysis of multimodal texts and their reception.

## 5 Materiality

Materiality has played a key role in multimodal research since the very first discussions of the theoretical and analytical preoccupations of this line of scientific enquiry. Kress and van Leeuwen (2001) highlight this very clearly:

A semiotics which is intended to be adequate for the description of the multimodal world will need to be conscious of forms of meaning-making which are founded as much on the physiology of humans as bodily beings, and on the meaning potentials of the materials drawn into culturally produced semiosis, as on humans as social actors. All aspects of materiality and all the modes deployed in a multimodal object/phenomenon/text contribute to meaning (p. 28).

However, before looking at how materiality affects semiosis, it is worth discussing *why* it is important for the material to be part of multimodal research. Bateman (2021a, p. 36) highlights the role materiality can play in providing "a robust empirical methodology for multimodality studies." As he argues,

Focusing attention on materiality naturally brings into close relief those very 'objects of analysis' (construed quite literally) that are of central concern for multimodality. It will consequently be argued that a better understanding of materiality contributes directly to methodology in that knowing more about materiality also supports more robust and well designed empirical studies (p. 36).

Bateman et al. (2017, p. 230) pose as the first step of empirical investigation the identification of the material properties of the communicative event under analysis. The four basic dimensions they discuss are temporality, space, role, and transience. Temporality refers to whether the traces are "dynamic" (e.g., a film) or "static" (e.g., a page in a book); space refers to whether they are two or three-dimensional; role refers to the relationship between an interpreter and the communicative event as being either 'observational' (i.e., as placed outside of the event) or "participatory" (i.e., as place inside of the event); transience refers to the traces being either "permanent" (e.g., ink on paper) or "fleeting" (e.g., gestures) (Bateman, 2021a, p. 40). Multimodal texts will not only present all these four dimensions at once, but also potentially have co-occurring bundlings of signifying material that belong to different dimensions; this will require the analyst to "slice" the communicative event in smaller analytical units, or "sub-slices", for more fine-grained and precise analyses (Bateman, 2021a, pp. 42–43).

As Bateman (2021a, p. 43) also notes, however, the initial analysis of the four dimensions by itself cannot adequately deconstruct complex multimodal artefacts, since "it is not the case that situations can be positioned with absolute freedom along each of the dimensions given." To this end (p. 43ff.), he proposes a more nuanced categorisation of the dimensions, which sees, for example, the *role* dimension being problematised by the fact that different interpreters may engage with the same materialities in different ways through their embodied perception, thus blurring the dichotomy observer/participant. Similarly, the dimension of *transience* is also broken down further by looking at aspects such as manner of (dis)appearance,

<sup>5</sup> More recently, Oja (2023) provides yet another understanding of *modalities*, using the term to refer to sensory modalities and arguing for a clear differentiation between semiotic modes and sensory modalities.

degree of granularity and time depth. Bateman (2021a, p. 47, *my emphasis*), therefore, points out that “th[is] characterization of materiality [...] is not that of physics but rather rests on *active perceivers’ embodied engagement* with materials for semiotic purposes.” This semiotically-oriented view of materiality expands empirical avenues for researching the *reception* of multimodal artefacts, since research in this area has highlighted how engagement differs depending on a number of individual factors, such as “the task or goal of the [text] examination, previous knowledge and expertise, expectations, emotions and attitudes. Apart from viewer characteristics, even the context in which [texts] are displayed, perceived and interpreted plays a role” (Holsanova, 2014, p. 340). This is an aspect I will discuss further in the next section.

Once the relevant slices and materialities are identified, one can proceed to analyse what different modes, and the canvases and media that support them, contribute to signification, and how they do so. At this level of analysis, one of the aspects often discussed in approaches to multimodal research is the idea of the *affordances* of materials. Affordances can be defined in terms of what the environment that surrounds us, whether of a natural or artificial type, allows us to do. As Gibson (2015, p. 120) puts it, “[t]he different substances of the environment have different affordances for nutrition and for manufacture. The different objects of the environment have different affordances for manipulation.” Amongst the objects that allow for manipulation, multimodal scholars have routinely included the modes and media of communication. Moreover, Gibson (2015, p. 121) argues that these affordances are neither subjective nor objective, or rather, that they can be both depending on the context and the observer/user; this is a property that can also be found in the Hallidayan concept of *meaning potential*, which has prompted some to see affordances as synonym of semiotic resources (Van Leeuwen, 2005, p. 5). The issue of what the affordances of modes and media are has certainly been the one discussed the most in the literature, either in terms of “abstract distinctions and commitments” (Bateman and Schmidt, 2012, p. 94) or in terms of their ideological load (Machin, 2013, pp. 349–350). However, Bezemer (2023, p. 6, *emphasis in original*) notes how Kress defines affordances “in terms of (i) *materiality* or *inherent* physical properties; and (ii) *social and cultural conventions* of using these properties for communication.” The second point hints to Gibson’s idea of affordances being both subjective and objective, but defines more clearly where the subjectivity lies, that is in the socio-cultural conventions surrounding communicative events and the modes deployed therein, which are inevitably contingent to historic specificities. This dual nature of affordances, therefore, points to different lines of enquiry that need to be taken into consideration when approaching multimodal communication.

On the one hand, we can focus on how the materiality of the signs and sign systems affect their deployment in communication. This can include looking at what semiotic resources can be accessed by specific semiotic codes and what material form they will take once deployed through specific media and genres as part of their mode contribution. Moreover, we can also look at how the materiality of the signs and their paradigmatic and syntagmatic structure bear representational force. On the other hand, we can focus on how socio-cultural conventions, as well as technological advancement, shape and alter the range of material configurations that can be deployed through specific modes, media and genres. Both lines of enquiry are pursued by Bateman (2014) when analysing the historical development of the

genre of “bird field guides”. Along the first line of enquiry, he considered “what semiotic modes are being mobilised in the service of what kinds of rhetorical strategies” (p. 252) in various reiterations of the same genre at different points in time and through both non-digital and digital media.<sup>6</sup> Along the second line of enquiry, he highlighted that, although the construction and deployment of rhetorical strategies relied on different modes as the media deployed changed over time unlocking new and different affordances to the modes, the rhetorical strategies themselves, as a communicative characteristic of the genre, remained unvaried.

It is worth noting at this stage, that both lines of empirical enquiry, which will be discussed in more detail in the next section, can be approached both from a formal, structuralist perspective and, as in this case of approaches within social semiotics and multimodal critical discourse analysis, from social and critical perspectives.

## 6 Implications for empirical research in multimodality

This section will look at the implications for empirical research in multimodality based on the discussion so far and will concentrate on the new conceptualisation of *semiotic resources* as carriers of metafunctional constructs as well as on the role of materiality as discussed in the previous section. Moreover, the discussion will cover both multimodal text analysis and multimodal text reception as well as pointing out aspects that can be of use to social and critical approaches to multimodal research.

The focus on metafunctions as a legitimate empirical avenue of research has recently been acknowledged by Bateman (2021b, p. 3), who points out that “[m]etafunctional accounts offer interpreters and producers resources for discussing and reflecting on just how information is structured and expressed and the social positions that appear to be being taken up in and by messages.” However, Bateman (2021b) also adds that:

Currently, descriptions such as those employing metafunctional distinctions [...] presuppose particular kinds of meanings for forms of expression a priori – that is, many current frameworks in use conflate the identification of technical features, i.e., identifiable material forms, and those features’ meanings [...] Reliably applicable categories have, however, not yet been established by corresponding empirical investigations of the semiotic resources considered. Establishing and developing a more reliable foundation for such descriptions therefore needs to be made a priority. (p. 4).

The categorisation of semiotic resources as proposed in this paper, if investigated both at the stages of multimodal text production *and* reception, can serve as a starting point to build those ‘reliably applicable categories’ Bateman calls for. The materiality of modes,

<sup>6</sup> Bateman (2021b) further debunks the idea that digital media are to be treated differently than traditional media, and provides a taxonomy of configurations that can be applied to all communicative events.



media and sensory channels<sup>7</sup> can provide research hypotheses based on their affordances and relation to semiotic resources. Hypotheses can then be tested empirically both from the perspective of text production and from the perspective of text reception.

As a starting point, three discrete, but related research focuses can be identified (cf. Bateman, 2021b, pp. 3–4): (i) investigating which semiotic resources and metafunctions individual modes can actualise (cf. also Bezemer, 2023, p. 11ff.); (ii) investigating the relationship between different modes actualising the same semiotic resources and metafunctions; (iii) investigating the contribution of individual modes to perform the three metafunctions of a communicative event. These research focuses can be pursued both qualitatively and quantitatively as neither paradigm is intrinsically better than the other in multimodal research, provide quality criteria are in place (Pflaeging et al., 2021, p. 6ff.). Moreover, these research focuses can be pursued from the perspective of both text analysis and text reception.

## 6.1 Investigating which semiotic resources and metafunctions individual modes can actualise

The first focus is on the material affordances of individual modes and the extent to which they can (and indeed do) actualise certain semiotic resources in a specific communicative context. The analysis would consider both the materiality of the modes themselves and the materiality of the media through which they are deployed. The aim here is to establish the signifying potentials of modes and the material aspects involved in the process of signification within specific communicative practices. Hypotheses can be generated from existing theory and qualitative studies and then tested on specific corpora, be these medium-based, genre-based, or a combination of both. Discrete sets of semiotic resources, *discursive*, *pragmatic*, *stylistic* and *textual* can be investigated so that reliable categories can be confirmed or rejected.

Interestingly, comparisons can be made between the same modes as deployed across different media and genres. This could shed some light on whether certain resources can be activated in all contexts; whether they are activated in a similar fashion or in different ways and, if the latter, what factors (both in terms of individual choices of the text producers and as imposed from the context of deployment of the modes) influence the syntagmatic and paradigmatic organisation of the mode as well as the resulting discourse semantics; whether their deployment has changed over time, as in Bateman's (2014) study of rhetorical strategies in the 'bird field guides' genre, and again, what factors might have contributed to such change.

From the point of view of reception studies, hypotheses regarding individual modes and metafunctional realisations that are generated through text analysis studies can be tested for identification, comprehension and interpretation. Alternatively, reception studies can be the starting point (perhaps through more qualitative

approaches) for such categorisations, which can then be empirically tested on corpora of texts or on larger cohorts of participants in order to highlight trends and patterns.

These perspectives can inform both formal, structural approaches to multimodal research and critical ones. Researching the contextual factors that lead to certain realisations of metafunctions, or indeed to certain modes not performing a metafunction at their disposal in specific contexts, can point at aspects of power dynamics between institutions and practitioners, between participants in the communicative event, and so on.

## 6.2 Investigating the relationship between different modes actualising the same semiotic resources and metafunctions

The second focus may result from the analysis of the first as it may turn out that more than one mode is contributing to performing the same metafunctional construct. This may be by contributing to actualise the same semiotic resource, or it may be by contributing to perform the same metafunction by actualising complementary semiotic resources. The aim here is therefore to investigate the relationship between different modes when co-deployed in a multimodal artefact. Again, discrete categorisations of the semiotic resources along the lines suggested in this paper allows one to focus on specific resources and material realisations, not only by looking at communicative events as a whole, but also at slices and sub-slices of materials and related canvases as suggested by Bateman (2021a, pp. 42–43).

Similarly to what discussed with the first focus, hypotheses can be generated from existing theory and qualitative studies, and tested empirically and quantitatively across corpora and comparable datasets. This approach would enable investigations into patterns of co-dependency between modes as they realise specific semiotic resources and metafunctions; analysis of variances of the co-dependencies identified across media and genres; factors affecting variance, driven by both bottom-up and top-down considerations; medium- and genre-specific historical developments of semiotic resources and metafunctions over time.

Receptions studies here can add valuable information at three different levels: studies integrating psychophysiological measures (e.g., using eye-tracking technology) can offer insights on matters such as attention, focus and 'reading' paths, in order to investigate, for example, whether certain modes are mostly relied upon in the recognition of certain semiotic resources or metafunctions. Experimental studies manipulating the multimodal output or relying on qualitative instruments such as think-aloud protocols and retrospective interviews can offer insights on recognition and comprehension of semiotic resources and metafunctions when these are deployed by modes individually or co-deployed: these can incorporate the exclusion of expected metafunctional realisations and/or the inclusion of unexpected ones. Finally, qualitative studies can offer insights on the interpretation of certain semiotic resources and metafunctions as well as an assessment of their effectiveness vis-à-vis their intended use and function.

From a critical perspective, this approach can shed light on matters of persuasion, manipulation, legitimation and argumentation, or any other pragmatic goals, by investigating how certain modes and

<sup>7</sup> Due to the limitation in space, I have not been able to provide an adequate treatment of *sensory channels* in this paper. However, there is already some work in this direction (e.g., Oja, 2023) and I am myself working on a contribution to this discussion.



related semiotic resources achieve their goals. Again, the critical variant for this research focus can rely both on text analysis and text reception studies.

### 6.3 Investigating the contribution of individual modes to perform the three metafunctions of a communicative event

The third focus is a step up from the second one, with the aim to provide a more holistic description of which metafunctions and semiotic resources are present in a specific communicative event. Assuming a communicative event necessarily relies on engaging the participants at all three metafunctional levels, i.e., ideational, interpersonal and textual, the aim of this line of enquiry is to identify which modes perform a specific metafunction, whilst taking into consideration the other variables already mentioned, i.e., media and genres, and their situational specificities and historic development. Once again, the classification of semiotic resources proposed in this paper would facilitate focussed investigations that can span across modes and, within communicative events, across slices and sub-slices of materials.

Most of what highlighted for the second focus in terms of formulating hypotheses applies to this third line of enquiry too. However, a further analytical focus with this third approach is the co-dependency and individual contributions of sub-slices and slices of semiotically-charged materials to the overall performance of semiotic resources and metafunctions, thus taking the level of analysis from within individual canvases to across multiple co-occurring canvases. As with the previous levels of analysis, reception studies can be integrated here, in which slices and sub-slices of material can be manipulated for experimental purposes, and more qualitatively driven studies can explore matters of comprehension, interpretation and effectiveness of different (sub-)slices configurations.

More generally, and valid for all the three levels of analysis discussed, the advantage of a clearly distinct and defined concept of semiotic resources and a more nuanced understanding of how materiality affects signification at different stages of text production, distribution and reception, allows one to zoom in on specific aspects of semiosis and to be able to approach the object of research both qualitatively and quantitatively.

## 7 Conclusion

The paper has engaged with a crucial issue in multimodal research, that is a lingering confusion (or disagreement) around some key concepts needed to research and write about multimodality. The definition and composition of mode as proposed by Bateman (2011, 2016) and Bateman and Schmidt (2012) has been used as the starting point to provide a clearer distinction between the concepts of *semiotic mode* and *semiotic resources*. It has been argued that there is an ontological difference between these two aspects of semiosis, with the former being a combination of material and abstract elements and the latter having no materiality of their own, but the ability to manifest themselves through different materialities. Indeed, the metafunctional

properties attributed to semiotic resources are often deployed and articulated not only through different materialities but also through the simultaneous co-occurrence of different modes and through accessing different sensory channels, which is another ontological difference between semiotic modes and semiotic resources. Semiotic resources have therefore been defined in this paper as *abstract, potential metafunctional constructs that can be realised through different materialities and/or semiotic codes*, and have been organised in four areas: discursive, pragmatic, stylistic and textual.

Once established the nature and ontological status of semiotic resources, the issue arose concerning their relationship with semiotic modes. The paper has argued that semiotic resources should be part of the constitutive elements of a mode and be placed at an intermediate stratum between the paradigmatic and syntagmatic axes of organisation (the second stratum in Bateman's model) and discourse semantics (the third stratum in Bateman's model), thus creating a four-stratal definition of a semiotic mode. This new stratum of semiotic resources is necessary to explain *how*, in the process of deploying a mode (i.e., in the process of semiosis), the material substrate is organised in specific paradigmatic and syntagmatic forms to allow certain (and not other possible) interpretations at the level of discourse semantics. It has been also argued that the choice of semiotic resources to be adopted can be guided both by bottom-up (i.e., text producer-driven) and top-down (i.e., context-driven) factors.

Moreover, since "we do not find 'free-floating' instances of semiotic modes" (Bateman, 2017, p. 168) the concept of media has also been discussed and a relationship of independence between media and modes established (Bateman, 2017, p. 172). With all the main concepts in place, and since materiality has been playing an increasingly important role in multimodal research to the point that focussing on it is necessary to provide "a robust empirical methodology for multimodality studies" (Bateman, 2021a, p. 36), the role of materiality in multimodal research has been discussed. Here two lines of enquiry have been identified as being 'unlocked' by a material approach to multimodality: the first concerns how the materiality of the signs and sign systems affects their deployment in communication; the second concerns how socio-cultural conventions, as well as technological advancement, shape and alter the range of material configurations that can be deployed through specific modes, media and genres. Finally, based on the new conceptualisation of semiotic modes and semiotic resources, and on the discussion around the role of materiality, implications have been outlined for empirical multimodal research and pointers offered as for potential research endeavours that can focus both on text production and text reception at three different levels of analysis (cf. Bateman, 2021b, pp. 3–4): (i) investigating which semiotic resources and metafunctions individual modes can actualise; (ii) investigating the relationship between different modes actualising the same semiotic resources and metafunctions; (iii) investigating the contribution of individual modes to perform the three metafunctions of a communicative event.

The work of John Bateman has paved the way towards a more systematic and empirically oriented way of doing multimodal research, especially within the social semiotics and SFL orientations. This paper is an attempt to build on this body of work and continue to strive for theoretical and methodological clarity in a discipline that is still in the process of establishing its own grounds and agreeing on

key concepts, despite the incredible body of research carried out over the past three decades.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

JC: Conceptualization, Writing – original draft.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. Research funding at Canterbury Christ Church University covered the APCs for this paper.

## References

- Bateman, J. A. (2011). "The decomposability of semiotic modes" in *Multimodal Studies: Multiple Approaches and Domains*. eds. K. L. O'Halloran and B. A. Smith (London: Routledge), 17–38.
- Bateman, J. A. (2014). "Genre in the age of multimodality: some conceptual refinements for practical analysis" in *Evolution in genre: Emergence, variation, multimodality*. eds. P. Evangelisti-Allori, V. K. Bhatia and J. A. Bateman (Frankfurt am Main: Peter Lang), 237–269.
- Bateman, J. A. (2016). "Methodological and theoretical issues in multimodality" in *Handbuch Sprache im Multimodalen Kontext*. eds. N. M. Klug and H. Stöckl (Berlin: Walter de Gruyter), 36–74.
- Bateman, J. A. (2017). Triangulating transmediality: a multimodal semiotic framework relating media, modes and genres. *Discourse Context Media* 20, 160–174. doi: 10.1016/j.dcm.2017.06.009
- Bateman, J. A. (2021a). "Dimensions of materiality" in *Empirical multimodality research: methods, evaluations, implications*. eds. J. Pflaeging, J. Wildfeuer and J. A. Bateman (Berlin: Walter de Gruyter), 35–63.
- Bateman, J. A. (2021b). What are digital media? *Discourse Context Media* 41:100502. doi: 10.1016/j.dcm.2021.100502
- Bateman, J. A., and Schmidt, K. H. (2012) *Multimodal film analysis: How films mean*. New York: Routledge.
- Bateman, J. A., and Wildfeuer, J. (2014). A multimodal discourse theory of visual narrative. *J. Pragmat.* 74, 180–208. doi: 10.1016/j.pragma.2014.10.001
- Bateman, J. A., Wildfeuer, J., and Hiippala, T. (2017) *Multimodality: Foundations, research and analysis: A problem-oriented introduction*. Berlin: Walter de Gruyter.
- Bezemer, J. (2023) What modes can and cannot do: *Affordance in Gunther Kress's theory of sign making*. *Text & Talk*. doi: 10.1515/text-2022-0055
- Bucher, H. J. (2017). "Understanding multimodal meaning making: theories of multimodality in the light of reception studies" in *New studies in multimodality: conceptual and methodological elaborations*. eds. O. Seizov and J. Wildfeuer (London & New York: Bloomsbury), 91–123.
- Chandler, D. (2017) *Semiotics: the basics (3rd ed.)*. London: Routledge.
- Elleström, L. (2010) The modalities of media: a model for understanding intermedial relations. In *Media Borders, multimodality and Intermediality*, ed. L. Elleström (Basingstoke: Palgrave Macmillan), 11–50.
- Forceville, C. (2009). "Non-verbal and multimodal metaphor in a cognitivist framework: agendas for research" in *Multimodal metaphor*. eds. C. Forceville and E. Urios-Aparisi (Berlin: Walter de Gruyter), 19–35.
- Gibson, J. J. (2015) *The ecological approach to visual perception*. New York & London: Psychology Press.
- Halliday, M. A. K. (1978) *Language as a social semiotic: The social interpretation of language and meaning*. Baltimore, MD: University Park Press.
- Hiippala, T., and Bateman, J. A. (2021). Semiotically-grounded distant view of diagrams: insights from two multimodal corpora. *Digit. Scholarsh. Humanit.* 37, 405–425. doi: 10.1093/llc/fqab063
- Hiippala, T., and Bateman, J. A. (2022). "Introducing the diagrammatic semiotic mode" in *Diagrammatic representation and inference: 13th international conference (diagrams 2022)*, Vol. 13462 of *lecture notes in computer science*. eds. V. Giardino, S. Linker, R. Burns, F. Bellucci, J.-M. Boucheix and P. Viana (Cham: Springer), 3–19.
- Holsanova, J. (2014). "In the eye of the beholder: visual communication from a recipient perspective" in *Visual Communication*. ed. D. Machin (Berlin: De Gruyter Mouton), 331–356.
- Kress, G. (2010) *Multimodality: A social semiotic approach to contemporary communication*. London: Routledge.
- Kress, G., and van Leeuwen, T. (1996) *Reading images: The grammar of visual design*. London: Psychology Press.
- Kress, G., and van Leeuwen, T. (2001) *Multimodal discourse: the modes and Media of Contemporary Communication*, London: Arnold.
- Machin, D. (2013). What is multimodal critical discourse studies? *Crit. Discourse Stud.* 10, 347–355. doi: 10.1080/17405904.2013.813770
- Machin, D. (2016). The need for a social and affordance-driven multimodal critical discourse studies. *Discourse Soc.* 27, 322–334. doi: 10.1177/0957926516630903
- Machin, D., and Mayr, A. (2012) *How to do critical discourse analysis*. London: Sage.
- Nord, C. (1991) *Text analysis in translation*. Amsterdam & Atlanta: Rodopi.
- O'Halloran, K. (2005) *Mathematical discourse: Language, symbolism and visual images*. London: Continuum.
- Oja, M. (2023). Semiotic mode and sensory modality in multimodal semiotics: recognizing difference and building complementarity between the terms. *Sign Syst. Stud.* 51, 604–637. doi: 10.12697/SSS.2023.51.3-4.05
- Page, R. (2009). "Introduction" in *New perspectives on narrative and multimodality*. ed. R. Page (London: Routledge), 1–13.
- Pflaeging, J., Wildfeuer, J., and Bateman, J. A. (2021). "Introduction" in *Empirical multimodality research: Methods, evaluations, implications*. eds. J. Pflaeging, J. Wildfeuer and J. A. Bateman (Berlin: Walter de Gruyter), 3–32.
- Stöckl, H. (2014). "Semiotic paradigms and multimodality" in *The Routledge handbook of multimodal analysis*. ed. C. Jewitt. 2nd edn (London: Routledge), 274–286.
- Van Leeuwen, T. (1999) *Speech, music, sound*. London: Macmillan
- Van Leeuwen, T. (2005) *Introducing social semiotics*. London: Routledge.

## Acknowledgments

I would like to thank Alex Cockain and John-Paul Riordan for their comments on the first draft of the paper. I would also like to thank the reviewers for their comments, which have considerably improved the clarity and focus of the paper.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



## OPEN ACCESS

## EDITED BY

Janina Wildfeuer,  
University of Groningen, Netherlands

## REVIEWED BY

Leandra Thiele,  
University of Bremen, Germany  
Heike Baldauf,  
Lumière University Lyon 2, France

## \*CORRESPONDENCE

Arianna Maiorani  
✉ a.maiorani@lboro.ac.uk

RECEIVED 03 January 2024

ACCEPTED 19 March 2024

PUBLISHED 28 March 2024

## CITATION

Maiorani A (2024) The materiality key: how work on empirical data can improve analytical models and theoretical frameworks for multimodal discourse analysis.  
*Front. Commun.* 9:1365145.  
doi: 10.3389/fcomm.2024.1365145

## COPYRIGHT

© 2024 Maiorani. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# The materiality key: how work on empirical data can improve analytical models and theoretical frameworks for multimodal discourse analysis

Arianna Maiorani\*

Communication and Media Department, School of Social Sciences and Humanities, Loughborough University, Loughborough, United Kingdom

This article is a critical reflection on the way the notion of materiality informed the project and the development of The Kinesemiotic Body project carried out by a UK and German research team and of the model of analysis it adopted, the Functional Grammar of Dance. It starts with an excursus of some of the most interesting developments in other discipline that turned to the investigation of materiality as an epistemological perspective, and it shows how the same type of focus has impacted multimodal discourse analysis focusing on movement-based communication. The overarching theme that characterises this multidisciplinary attention to materiality is its anchoring function to the temporal and spatial coordinates in which social phenomena are contextualised, which is taken as the fundamental condition for shaping our perception and understanding of the world in all areas of experience and knowledge. A more specific example of how the notion of materiality impacted the development of movement-based discourse analysis will be provided by an example of analysis of rich movement data captured live from professional dancers from the English National Ballet.

## KEYWORDS

materiality, multimodal discourse analysis, Functional Grammar of Dance, empirical data analysis, movement-based communication

## Introduction

This paper provides a critical reflection on the role played by the notion of materiality in the development of movement-based discourse analysis within the wider area of Multimodality studies. It is positioned within an even wider area of multidisciplinary research that focused on this notion in the last few decades and that foregrounded some very interesting points for reflection and development across disciplines. Through examples drawn from a recent research project in movement-based communication, it will demonstrate how in order to incorporate effectively the awareness and understanding of materiality in a communicative environment, it is essential to turn to the analysis of empirical data, which in turn provides solid evidence to strengthen and/or advance theoretical frameworks. The project in question is The Kinesemiotic Body, funded by the Arts and Humanities Research Council (AHRC) in the UK and the German Research Foundation (DFG) in Germany. The fact that the project focused on movement-based analysis (specifically on dance choreography) carried out by scholars from very different disciplines, where the importance of incorporating the materiality of the human body in interaction with a performance environment was considered through different

approaches (Multimodal Discourse Analysis, Engineering, Computer Science, etc.) makes the examples of empirical data analysis proposed here particularly appropriate to the consideration of the notion of materiality as an interdisciplinary one and provides a clear connection with John Bateman's discussion of materiality in relation to the development of Multimodality as a practice that encompasses borders between disciplines and research areas (Bateman et al., 2017; Bateman, 2019, 2022). This article will also show how the consideration of the materiality of dance allowed *in primis* for the further development of the Functional Grammar of Dance (Maiorani, 2021; Maiorani et al., 2022; Maiorani and Liu, 2023), which is now a more comprehensive and even more flexible tool that scholars have started to use for analysing movement-based communication in dance performances other than ballet or even outside the domain of dance altogether (see Mouard Ruiz, 2021; Bolens, 2022; Meissl et al., 2022; Prové, 2022; Sindoni, 2022; Vidal Claramonte, 2022; Wu, 2022; Elyamany, 2023). The examples of empirical data analysis will be preceded by a presentation of how the Functional Grammar of Dance is implemented in ELAN, a widespread commercial, free-to-use software traditionally used for annotating conversations or verbal interactions, for which we created a complete set of interdependent tiers and controlled vocabulary. By including spatial annotation categories and the distinction of internal discourse structures, our annotation offers quite innovative insights into the way movement-based communication can be annotated and analysed.

In order to describe the impact of the concept of materiality on The Kinesemiotic Body project—and especially the way materiality was foregrounded by Bateman's work in multimodal discourse analysis—I need to take a series of steps backwards, to the time when Kinesemiotics, a new interdisciplinary research area, was developed at Loughborough University. Kinesemiotics started with an interdisciplinary team of researchers created at Loughborough University in 2016, where we covered Linguistics, Semiotics, Multimodality, sensor Engineering, and Computer Science. After receiving funding from the Loughborough University CALIBRE programme in 2017 to work in collaboration with the English National on the investigation of movement-based discourse analysis by capturing a small amount of dance movement data, we joined forces with John Bateman for a joint grant application to the Arts and Humanities Research Council (AHRC) and the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) and we were funded for the collaborative international project called The Kinesemiotic Body.<sup>1</sup> The project aim was to advance the understanding of movement-based communication starting from choreographed movement in a worldwide renown movement-based form of performance: ballet. This choice was driven by the team's specific expertise as well as by the pre-existing collaboration with the English National Ballet and the status of ballet as a form of performance based on movement with a tradition long recognised and established at

international level, a tradition that has had an enormous impact on the elaboration and the development of many other forms of movement-based performances. Our intention was to evaluate whether and how we could apply a linguistically-motivated model for the analysis of verbal discourse to the study of how the body communicates by interacting with the space within the context of a performance setup. This would not only allow to deepen our understanding of the specific form of performance on which we were focusing, but also to develop a new approach to non-verbal communication with a more finely elaborated notion of movement-based discourse structure.

To contextualise the results of the project within a much more comprehensive scientific overview, this article will start by considering how the same notion has been approached by different disciplines in recent years looking at some significant examples of literature, thus showing how this concept actually taps into the very foundations of a multidisciplinary idea of knowledge.

## Materiality across research areas and disciplines

In several research areas, materiality seems to be considered as a contextualised configuration of the spatial and temporal location of multimodal communication, an architecture that anchors theory to real-life situations and allows for the encounter and cross-fertilisation of diverse fields of study. Contemporary ontological philosophy puts materiality at the centre of social life and interaction (Schatzki, 2010), positing social phenomena as configurations of practices and material arrangements, thus recognising materiality itself as a component of social phenomena that combines with technology and practices. In this way, the relationship between practices (including meaning-making practices) and the material arrangements in which these practices take place spatially and temporally – the fundamental socio-cultural coordinates – becomes the focus of contemporary social ontology. Schatzki (2010, p. 125) also points out that materiality is not merely physicality: it is rather to be defined as 'composition', the 'stuff' of which social life is made. The issue at stake is therefore to find a way to describe it systematically. Schatzki (2010, p. 129) also defines practices as 'organized spatial-temporal manifolds of human activity. Examples are cooking practices, political practices, manufacturing practices, football practices, dating practices, and horse breeding practices'. The material arrangements that form nexi with practices to generate social phenomena are 'sets of interconnected material entities' (Schatzki, 2010) that can be human beings, artifacts, organisms, and other natural items. The materiality of a social phenomenon can therefore be extremely complex, and the problem is finding a systematic way to pick up the elements that compose it.

In the area of semiotic studies, and essentially drawing on Peirce, the complexity and centrality of the notion of materiality has already been foregrounded by Petrilli (2008) through the specification of two types of materiality that inhere the sign itself: 'In a global semiotic perspective, it would seem that the first claim to be made is that the existence of biological material is the initial condition for sign material or semiosis material to exist. It goes without saying that no less necessary for the existence of biological material is the existence of chemico-physical material. Therefore, we could begin by stating that the materiality of signs presents itself on various levels, upon which

<sup>1</sup> The collaboration with John Bateman and the University of Bremen had actually started with John's fellowship at Loughborough University funded by the Institute of Advanced Studies and aimed at fostering our interdisciplinary collaboration in 2017. The rest of the research team was made by Massimiliano Zecca, Russell Lock, Chun Liu from Loughborough University and Dayana Markhabayeva from the University of Bremen.



basis we may propose a typology of semiotic materiality' (Petrilli, 2008, p. 139). Understanding semiotic materiality in this respect involves the recognition of a clear distinction between physical and biological materiality, the latter generating a further distinction between living and non-living organic materiality. Materiality is therefore seen as being at the origin of human experience of communication.

One of the most interesting examples of the analysis and use of materiality as a foundational epistemological concept comes from energy studies, where energy is conceptualised in its materiality to understand how its perception impacts on daily practices and transactions worldwide. A whole trend of energy studies has been working for decades on the reconceptualisation of the very notion of 'energy' by drawing on multidisciplinary and interdisciplinary approaches that include also theories from geography, politics, history, anthropology, etc. All these approaches focus on the effort to define the *materiality* of energy. Balmaceda et al. (2019) pose four fundamental questions to open a fruitful dialogue and exchange amongst different research areas; their queries are about the *location* of energy materiality, its users and the way they use it, its relational characteristics with context in terms of spatial and temporal scales, and the analytical role of energy materiality in the different epistemological areas. These questions are meant to anchor a theoretical enquiry on this fundamental notion to specific contexts in real life, such as the way energy materiality determines constraints in agency that will then impact on infrastructures and politics (i.e., energy consumption, supply chains, etc.). These questions also highlight the historical relationship between energy materiality and the evolution of technology (Leonardi and Barley, 2010), which is also a factor that impacts in a fundamental way social semiotics practices across time and space. Dance discourse—the meaning produced and shaped by choreographed movement in dance performances—is movement-based and movement involves the flow of kinetic energy. The way we experience and capture this flow for various purposes (archiving, documenting, visualising, etc.) is also impacted by the development of technology and of the devices that allow us to anchor to a specific time and place performances that would otherwise be lost once they have taken place. To understand how these questions may be of considerable relevance even when studying the development and perception of dance discourse, it suffices to think of the way an audience perception of live dance performances has changed considerably during and after COVID 19 lockdowns, when the perception of temporal and spatial *location* of performances worldwide was dramatically changed by the impossibility of actually attending a live performance in theatres. It was the audience's perception of these coordinates that technological affordances successfully managed to change when a number of theatre and ballet companies survived thanks to the broadcast of performances originally recorded for live streaming in cinemas and then turned into 'live pre-recorded events' packaged for home entertainment (Maiorani, 2020).

The importance of how the flow of movement is anchored to a spatially and temporally located context also emerges in trans-contextual analysis, a branch of social semiotics that looks at how materiality is perceived in different contexts through mobility. Kell (2015, p. 425) proposes the concept of 'meaning-making trajectories which are made up of recontextualizing and resemiotising moves'. This concept is meant to incorporate the flow of meaning movement and

transformation within contextualised communication through language, and it is linked to the materiality of communicative contexts moving across time and space; it also resonates with that of *trajectory* in Minimal Ballet Sequences, a unit of dance discourse analysis that I will explain below and that provides the description of dance discourse with a connective thread that incorporates the flow of movement and allows for the understanding of the different functions of *orientation* and *direction* in movement-based discourse. Meaning-making practices in trans-contextual analysis do not only take into consideration movement across contexts but also the role of material objects that interact with the 'text-artefacts' (Kell, 2015, p. 426), thus advocating for multimodality as a more comprehensive approach to the analysis of communication.

The connection between materiality, flow of experience and energy and embodiment is also at the centre of several cutting-edge theoretical approaches to knowledge understanding in the humanities. Whilst creative writing practices and cultural anthropology interrogate the relationship between identity and the materiality of the semiotic forms (Wilf, 2011), experimental literature focuses on the notion of materiality when trying to provide a flow of multimodal experiences to its readers (Lee, 2014). The consideration of materiality becomes particularly crucial in translation practices, where the materiality of the text emerges in all its complexity, ranging from its physical features to the way the written word conveys auditory, tactile, visual, and other sensorially-perceived (in other words, multimodal) meanings. In this case and drawing on Gibbons's (2012) idea of reading as an activity involving multisensory perception, the embodiment of a text materiality is once more conceived as the anchoring of the reader's meaning perception of multimodal, multisensorial meaning to a specifically located spatial and temporal context. In this way, the reader's body and its physical environment, its way of perceiving the world through the senses, becomes the *nexus*, the filter through which the very act of reading, of perceiving the materiality of a text turns into its embodiment.

Whilst experimental literature focuses on the nexus between narrated spaces and topographies and the way these are perceived through reading, recent studies on national mobility and infrastructures also pay attention to the materiality of the environment and how it influences the emergence and understanding of nationalisms and national identities (Merriman and Jones, 2017). Also in this case, materiality is theorised as the constellation of materials of diverse nature that anchors the flow of multimodal discourse—one of nation and identity—to specific temporal and spatial locations or to the process of crossing them. This perspective was generated by a wider context of studies on the relationship between discourse and materiality and its impact on management and organisational theory (Putnam, 2014), which has at its centre the dimensions of time and space and sees communication as the *locus* of the interplay between human agency and discourse.

Educational contexts have also turned to the study of the materiality in the context of traditional teaching and learning activities to develop more updated and effective pedagogical strategies within the perspective of multimodality. The materiality of multimodal forms of feedback has been studied to improve and update current forms of teaching and learning practices (Tyler, 2021), whilst lectures have been considered as a form of 'multimodal, sociomaterial performance' (Lacković and Popova, 2021) that has the human body and movement-based communication at its centre. This



reconceptualisation of lectures as a multimodal, movement-based practice draws on the concept of *sociomateriality* (Gherardi, 2017), which is grounded in post-humanist studies and essentially describes the interplay between social structures and material contexts made of bodies and items interacting in space in which every day meaning-making practices are habitually carried out. This new epistemological approach to knowledge shuns from human-centred approaches to learning and considers human experiences through materially contextualised phenomena. One of its central areas of research is embodied work practices, which posits the human body as an epistemological focus.

In the more specific area of science education, the educational environment is seen as a synthesis of semiotic agents that interact to produce meaning (Pantidos et al., 2010). The teaching of physics is particularly seen as an activity that involves creating connections amongst different signifying items and anchoring them to specific spatial and temporal contexts to explain theories. This activity generates narratives that make use of verbal language, gestures, objects, graphs, body movement, etc., a specific teaching practice whose general features can be observed in all types of science teaching. In this respect, science teaching is very similar to theatre practice, and its materiality is very similar to the materiality of theatre, where narrative spaces are characterised by referents whose meanings define a specific semiotic landscape anchored to a specific time and space (which is more or less what happens with the set-up of a dance performance space). The materiality of these narrative spaces is also similar to those used to teach robots when providing them with exemplary situations: thus, concepts are taught and learnt by anchoring them to the spatial and temporal materiality of a real-life context, to the materiality of everyday semiosis that is shaped into meaning through discourse (Björkvall and Karlsson, 2011). Björkvall and Karlsson draw strongly on social semiotics and anchor the specificity of contextual materiality into culture: according to them, materiality offers a meaning potential (Kress, 2010) that is then shaped through meaning making practices grounded in specific cultures. The shaping activity of cultures also involves choices amongst affordances that will be selected to become semiotic resources for communication. As a matter of fact, in a specific context within a specific culture, not all material affordances will become semiotic affordances. It is therefore the regularity and recurrence of configurations of semiotic affordances that allows us to identify *modes* (Bateman et al., 2017) within specific temporally and spatially located cultures: ‘for an affordance to be turned into a semiotic resource, it needs to be picked up by a culture or by a social group and be continuously worked upon in activity types of various kinds. In other words, the affordance needs to be shaped by culture to become what we call a semiotic resource. From this it follows that even if affordances are material resources for humans to perceive when they act in their environment, they are not necessarily semiotic resources. However, also affordances that are not defined as semiotic resources can have meaning potential’ (Björkvall and Karlsson, 2011, p. 147).

As it will be demonstrated below, the challenge of understanding which affordances in dance are regularly and consistently used as semiotic resources was one that was faced by The Kinesemiotic Body project and one that benefited from the consideration of the notion of materiality as an external language for description, a language that applies to the analysis of rich, live-captured movement

data by taking into consideration the specific configurations of materials that are shaped into semiotic resources in the meaning-making practice of dance performances.

## A theoretical framework to anchor the flow of dance to its materiality

The theoretical framework of The Kinesemiotic Body project was strongly based in linguistic theory and multimodal analysis; besides the Functional Grammar of Dance (FGD, Maiorani, 2017, 2021), our work also drew on segmented discourse representation theory (Asher and Lascarides, 2003; Bateman and Wildfeuer, 2014) as well as on recent developments in corpora analysis, live movement data collection and data visualisation. Working both on video materials and on movement data collected live from professional dancers of the English National Ballet (ENB)—who performed whole ballet sequences both as single performers and in couple—we developed a method of multimodal annotation using the annotation software ELAN that allowed us to annotate and analyse not just how movement is structurally carried out along a temporal line and within a specific space, but also how through structured movement sequences, dancers communicate by projecting their body parts towards meaningful portions of space, thus creating semantic connections that guide the audience’s interpretation. In this way, we created a method for annotating dance sequences that incorporates both movement structures and meaning structures in a flow of data. To show the effect of the research carried out through the analysis of empirical data within The Kinesemiotic Body project, I will first describe the original version of the Functional Grammar of Dance model and then I will introduce the updated version with all the relevant additions.

## The first model of the Functional Grammar of Dance

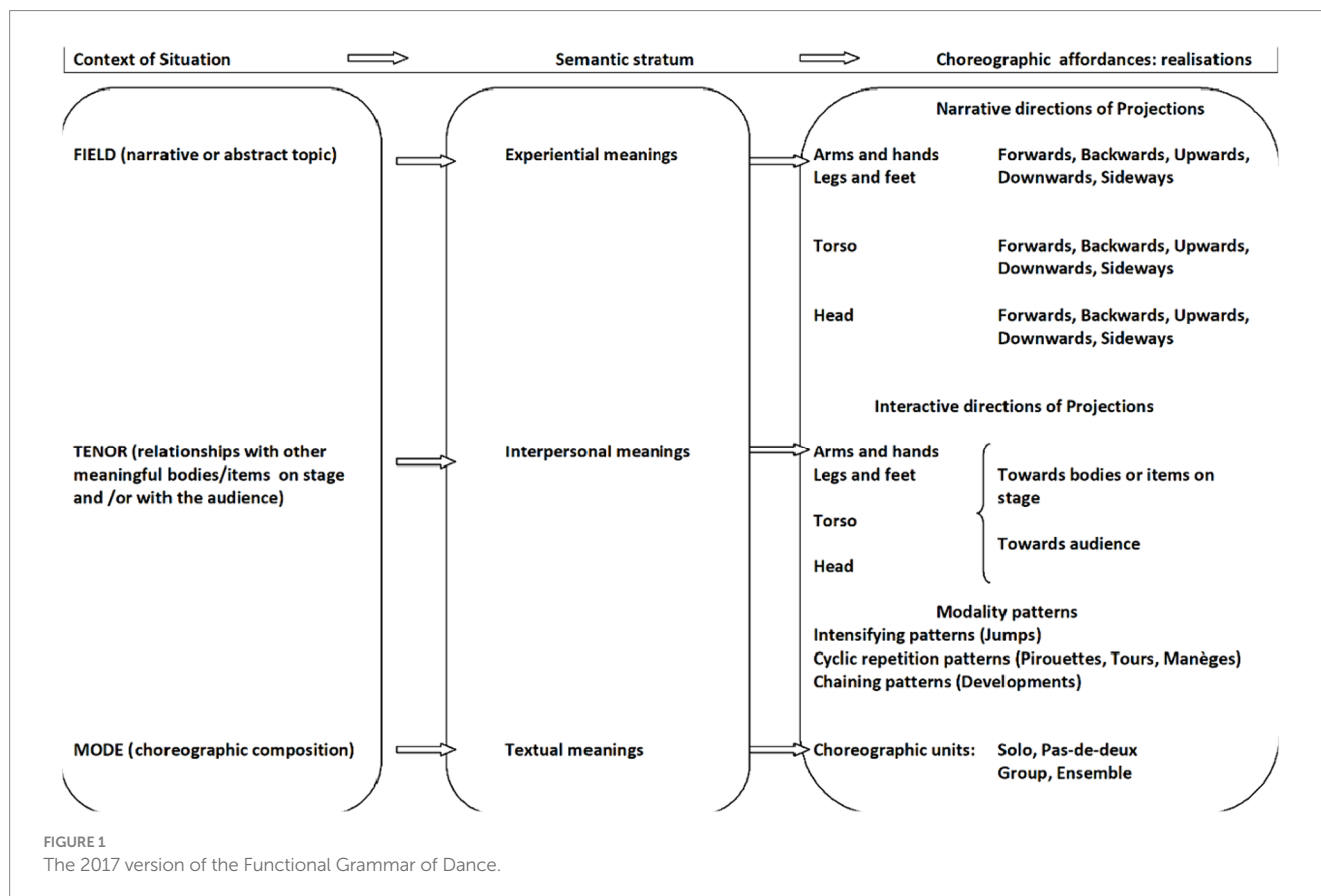
The first model of the Functional Grammar of Dance was published in 2017 and it clearly drew on Halliday’s Functional Grammar for verbal language. The model was already completely different from traditional dance notation systems as it was created and used for the manual analysis of dance discourse (movement structures *and* corresponding meaning), not for the notation of deconstructed movements and their physical qualities. With respect to the current and updated FGD model, it was simpler but it already incorporated first and foremost the dancer’s point of view as the starting point of movement, even if at the time it had only been applied to manual analysis of video clips. The FGD posits that movement-based communication, like verbal communication, always happens in a specific Context of Situation whose variables, Field (what is happening), Tenor (who is taking part), and Mode (how communication is being carried out by the participants to the communicative event), activate as many meanings (respectively Experiential, Interpersonal, and Textual) that will then be realised by different linguistic structures. In movement-based communication, these structures are also movement-based and they are called Choreographic Affordances, namely all possible body-structure combinations performed by dancers whilst moving,

structures that vary according to the dance style that is being adopted and the possibilities and limitations of human bodies. Figure 1 shows the first version of the FGD.

In order to generate meaning, choreographic affordances allow for the creation of structured *projections* of body parts towards meaningful portions of the performance space. The first fundamental difference between the Hallidayan Functional Grammar and the FGD is the distinction between *physical space* and *contextual space*. Space is a fundamental dimension of the FGD: “A body is a spatial construct. It exists and functions through its relationship with space. Space itself is defined by the presence of bodies of any kind: without bodies, we call it ‘void’” (Maiorani, 2021, p. 1). From its first version, the FGD posited that whereas dance movement can be *instantiated* during training classes by dancers carrying out choreographic combinations in a studio’s physical space just for the purpose of training, meaning is only created through the interaction between body structures and the performance space, which is populated by contextually relevant objects, people, props, etc., and it is therefore *designed* for this purpose. Whilst dancing a choreography, dancers extend in various manners their body parts towards meaningful spots in the performance space, thus creating interactions between their dancing bodies and people, or objects, or props, or light effects, and these interactions will provide the audience with cues to follow a narrative, to understand who is interacting with whom or what, and to enjoy the choreographed sequences as a whole. The visualisations of these interactions are called *Projections*: the narrative ones indicate

action (i.e., going to, coming from, locating, connecting, addressing, engaging, etc.), the interactive ones indicate interaction either with the audience (AU) or with participants on stage (POS). Only in the contextual space—whether actually built or just imagined during rehearsals it does not matter, provided that there is a shared awareness of it—can dance discourse actually be realised.

Interestingly, the elaboration of the FGD also allowed for a more in-depth discussion of the discussion of *instantiation* as a foundational concept of Systemic Functional Linguistic theory, leading to its definition as a dynamic relationship and to a further elaboration of the theory of Context (see Maiorani and Wegener, 2022). However, the first FGD model in Figure 1 shows how some areas of analysis could not be fully developed without the use of a larger amount of data collected live from dancers: the whole model is based on the development of its theoretical foundations and on manual analysis performed on small scale video data and drawing on a solid knowledge of the range offered by choreographic affordances, especially in terms of ballet. The lack of analysis of richer data sourced from different dancers performing different roles shows particularly in the area of Textual meanings, which was still developed on merely theoretical assumptions that needed to be tested empirically. The work carried out through The Kinesemiotic Body project on a corpus of live-dance captured data provided exactly this opportunity to test the FGD application empirically and to develop an analytical method that could be implemented in a widely commercially available software for annotation.



## The updated model of the Functional Grammar of Dance and our annotation system

The updated version of the Functional Grammar of Dance was elaborated whilst annotating rich live-captured movement data with the ELAN software. The annotation system we have developed is not an alternative to traditional notation systems like Labanotation or Benesh notation, which involve intensive training in using specific scores and symbols and provide a notation of the physical characteristics and qualities of unstructured movement along the music score. These systems are movement notation systems. The FGD annotation we implemented using ELAN is a dance discourse annotation that always puts the dancer's point of view at the centre of each movement and provides information both on movement structures and on discursive structures using labels that make no use of specialistic terminology. The FGD annotation is a dance annotation method, which implies that dance is not considered only as physical movement but as a meaningful and contextualised movement-based performance (Maiorani, 2021; Maiorani et al., 2022). Our annotation system in ELAN develops on different levels: the lower level of Move, which is the basic unit of analysis of the FGD, and the level of Minimal Ballet Sequence (MBS), which is the smallest discursive unit and comprises two consecutive Moves. The annotation is based on the work of the body articulators: head, torso, arms and hands, legs and feet. The Move marks the minimum movement across space performed by a dancer and is delimited by a starting set of projections and an arrival set of projections. The two consecutive Moves in an MBS are the smallest discursive unit that provides a *trajectory* in direction: if the Move direction is the same for both consecutive Moves, the MBS trajectory is defined as *continuous*; if the direction changes, the MBS trajectory is defined as *varied*. When the choreography requires it, we also annotate at the level of Elaborations: these are extra arrival sets of projections that mark a change in position of the body articulators at the end of a Move that does not involve any movement across space.

The use of sensors and the related software to capture live movement data from dancers immediately showed us that we had to deal with a complexity of movement parameters that needed to be ordered and put into clear functional relationships in the annotation. As soon as we started working with ELAN, we realised that we had to make three dimensions of annotation visibly distinguished and integrated at the same time: the level of *physical movement*, which accounts for the way each body articulator moves in relation to the surrounding space coordinates (i.e., inwards/outwards, up/down, backwards/forwards, etc.), the level of *structure*, which accounts for the way the different body articulators are positioned with respect to the Move direction, and the level of *projections*, which accounts for the narrative and interactive values of body parts *projections* towards meaningful portions of the performance space. These distinctions were necessary to show the complexity of the movement-based discourse enacted by dance, where the meaning created by projections is determined also by the position of articulators with respect to the immediate space references and the direction that a whole Move has taken. This complexity of relationships became visible when we started capturing live-data from dancers and had to take into consideration all the elements of movement we had to measure in order to account

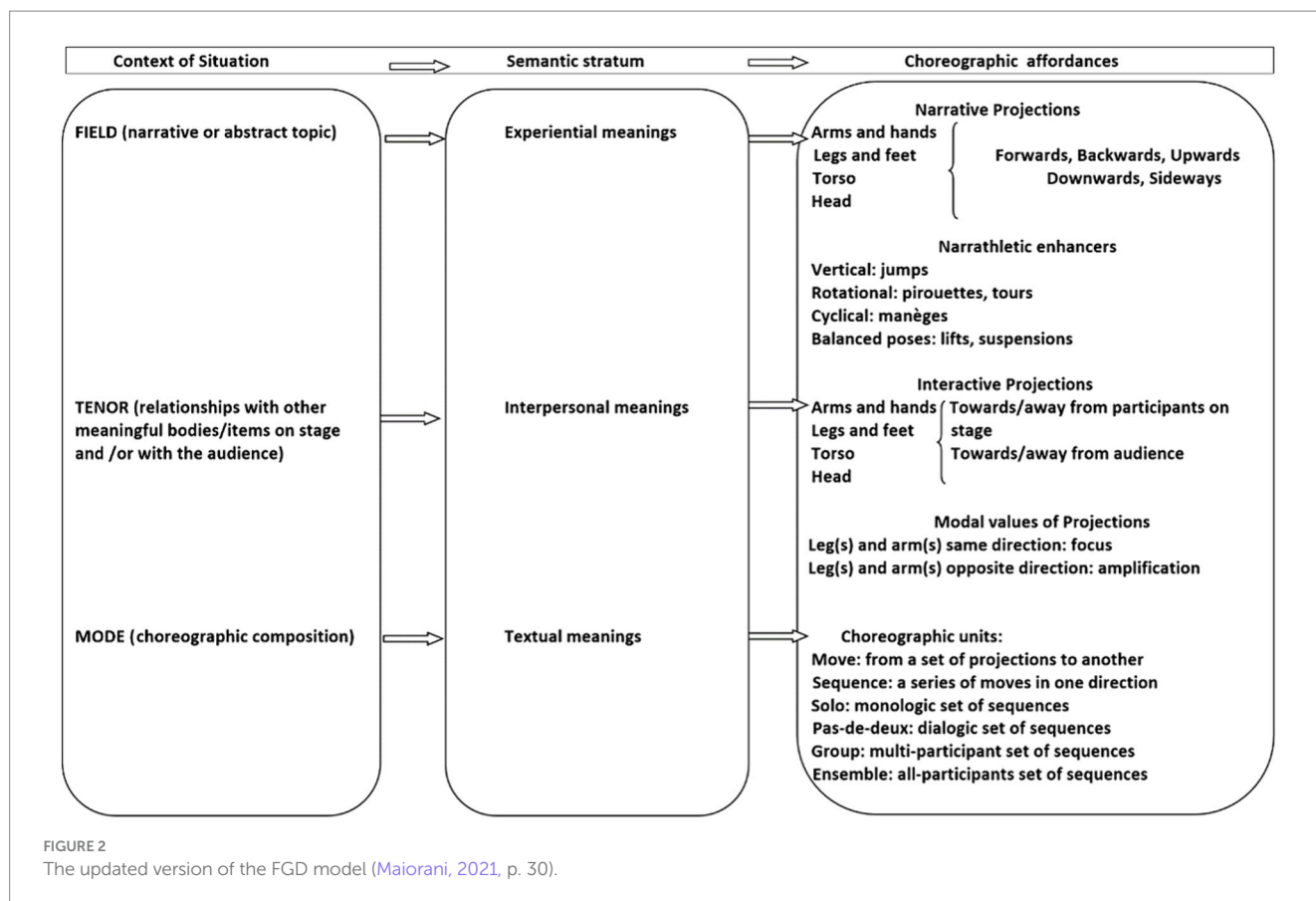
for all the factors that determined choreographic choices. The tiers related to each dimension are separated and colour coded but they are all at the same level, thus allowing for the visualisation of the complexity of factors all contributing at the same time to the realisation of projections within the performance space. After segmenting the flow of data into Moves, the first tier we annotate is always that of physical movement, which provides us with a picture of where every articulator is at the moment of annotation with respect to the immediate spatial references as they are perceived by the dancer; then we annotate the structures, which incorporate the direction the dancer takes when moving and the respective positioning in space of all the articulators with respect to movement; finally, we incorporate the discursive dimension by annotating narrative and interactive projections, which shows what type of actions and interactions the choices made in terms of physical movement and structure determine. The version of the FGD we used within ELAN is the most recent one, which we started developing after a preliminary work of live-movement capture data with the English National Ballet in 2017 and then kept on elaborating during The Kinesemiotic Body project. The impact of this work carried out on empirical data is reflected in a more detailed distinction of units of analysis (Choreographic units) specifically devised for empirical data segmentation and in the inclusion of narrathletic enhancers (showcasing dancers' athletic qualities) and modal values of projections (highlighting concentrations of projections in one direction) that reinforce the integration of physical and semantic description of the collected data. Figure 2 shows the current, updated model of the FGD (Maiorani, 2021, p. 30).

The improved work on the role of Move direction carried out on empirical data also allowed us to understand the discursive role of *trajectories*, designed by two consecutive Moves, thus highlighting the importance of segmenting MBSs. Figure 3 outlines the annotation framework we have developed.

When we transfer this annotation framework to ELAN, we create different tiers to annotate the Moves and MBSs. The highest level of description is that of the MBS tier, under which we annotate the Move tier (second description level) and Elaboration tier (third description level). The tiers with the descriptions of physical movement, structures, narrative projection, interactive projections, narrathletic enhancers and modal values of projections are all dependent on the Move tier and on the Elaboration tier when this occurs.

Every dance sequence is segmented according to the three levels of Move and couples of consecutive Moves are then grouped into MBSs. Therefore, the tiers depending on the Move (and Elaboration when present) align with Move (and Elaboration) segmentation. The end point of each Move aligns with the start of the subsequent one, and the same happens with the MBS segmentation, thus incorporating the flow of movement into the annotation. Figure 4 shows the FGD annotation framework implemented in ELAN.

The annotation tiers are linked to a controlled vocabulary divided into menus that provides specific options for each type of information annotated (i.e., physical movement, structures, projections, etc.) and draws on the FGD. The vocabulary is generated into a drop-down list in the ELAN annotation template, from which the annotator can select the most appropriate choice. The vocabulary does not contain any technical term and it is therefore very user friendly and open to non-specialist users. For



this reason, it can also be easily adapted to the annotation of movements other than ballet. Figure 5 shows an example of the drop-down controlled vocabulary list with options provided by labels that do not contain specialistic language.

The annotation template includes a window where the dance sequence that is being annotated is visualised and this can be reduced or enlarged in size according to necessity. Figure 6 provides an example of annotation made on data collected live from a dancer from the English National Ballet whose body was synthesised into an avatar. The figure does not offer the whole annotation but just a screenshot of a section as for the whole script it is necessary to scroll the text down. The video window has been reduced in size to provide a larger view of the annotation.

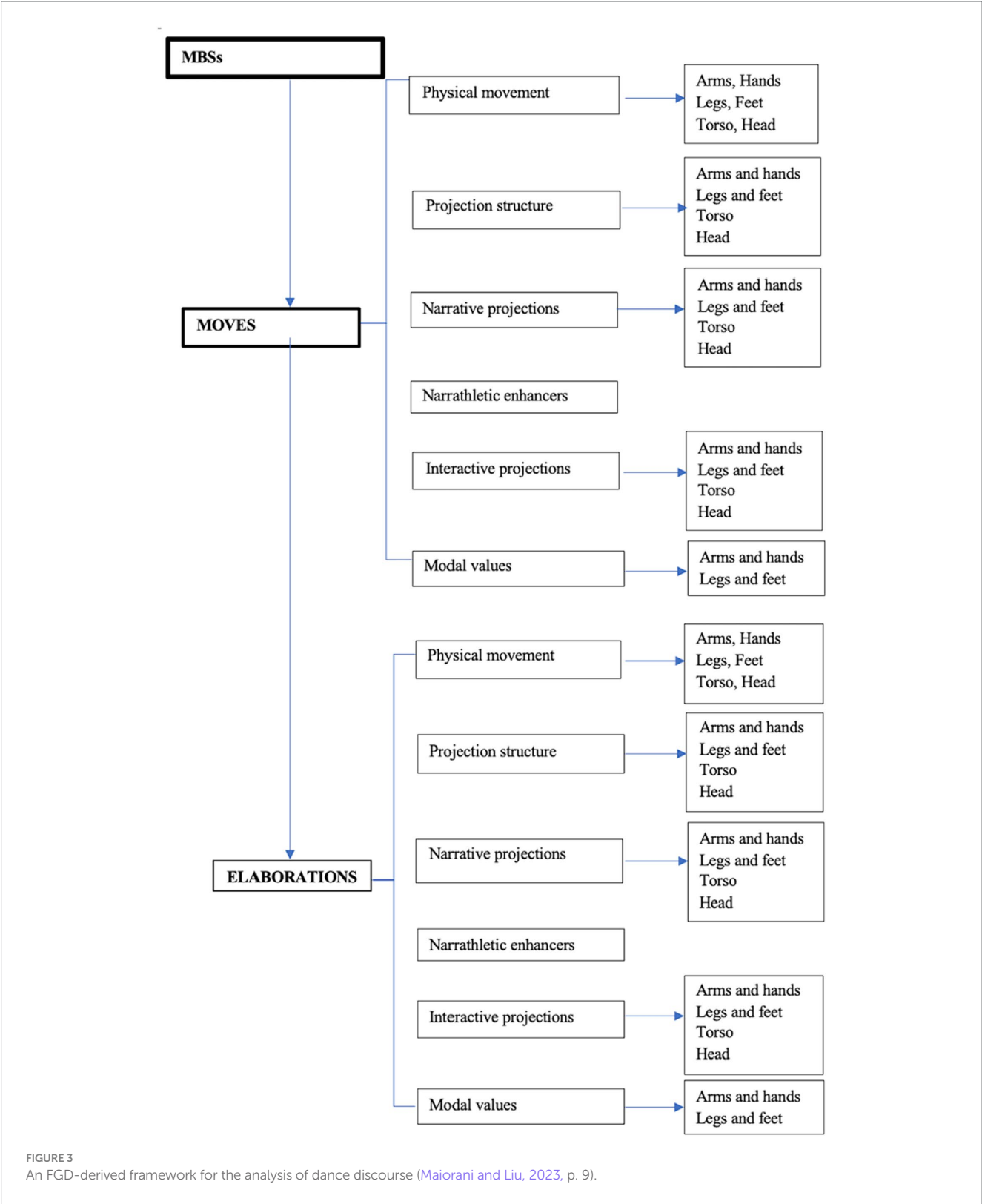
Figure 7 offers an example of annotation that highlights the segmentation into Moves and MBSs. The visualisation of the sequence that is being annotated is a video taken in a rehearsal studio at the English National Ballet headquarters in London.

The various tiers in which the annotation is organised is evidence in itself of the complexity of the materiality of dance that we were capable of capturing when working empirically and with live-captured movement data. The empirical work we carried out within The Kinesemiotic Body project allowed us to capture not only the relationships between movement structures and projections at various levels but it also made us realise that there are different levels of meaning carried out at different levels of discourse segmentation, and that *direction* and *orientation* have different and complementary roles in the perception of dance discourse, as will be discussed in the following section.

## Results of working with empirical data of a ballet sequence corpus

The FGD model was elaborated further when we started capturing live data from the dancers in a real rehearsal studio in preliminary work carried out in 2017 and then implemented in the analysis of the dance data corpus carried out with The Kinesemiotic Body project. When having to organise and annotate the data we recorded from dancers in rehearsal studios, the research activity based on data analysis had to face two main challenges. The first challenge was posed by the complexity of data which involved not only the dancers' movements but also the space set-up and the use of direction and orientation. Unlike what we had to take into consideration in the first examples of analysis performed manually with the FGD, where selected movement structures and projections were analysed on the basis of the systemic functional theoretical framework, a much greater amount of features and levels of communication deployment was suddenly available for analysis through the corpus of dance sequences collected with the English National Ballet. The second, consequential challenge was that the organisation of all these new features and levels that had not yet been captured or addressed by manual analysis had to be systematised in a consistent and replicable framework for annotation to be used for all items of the corpus. The initial manual analysis with the FGD had paved the road for a systematic investigation of dance discourse as movement-based communication in context but had not yet benefited from the amount of information provided by live-captured data. It lacked empirical application and was therefore limited in its

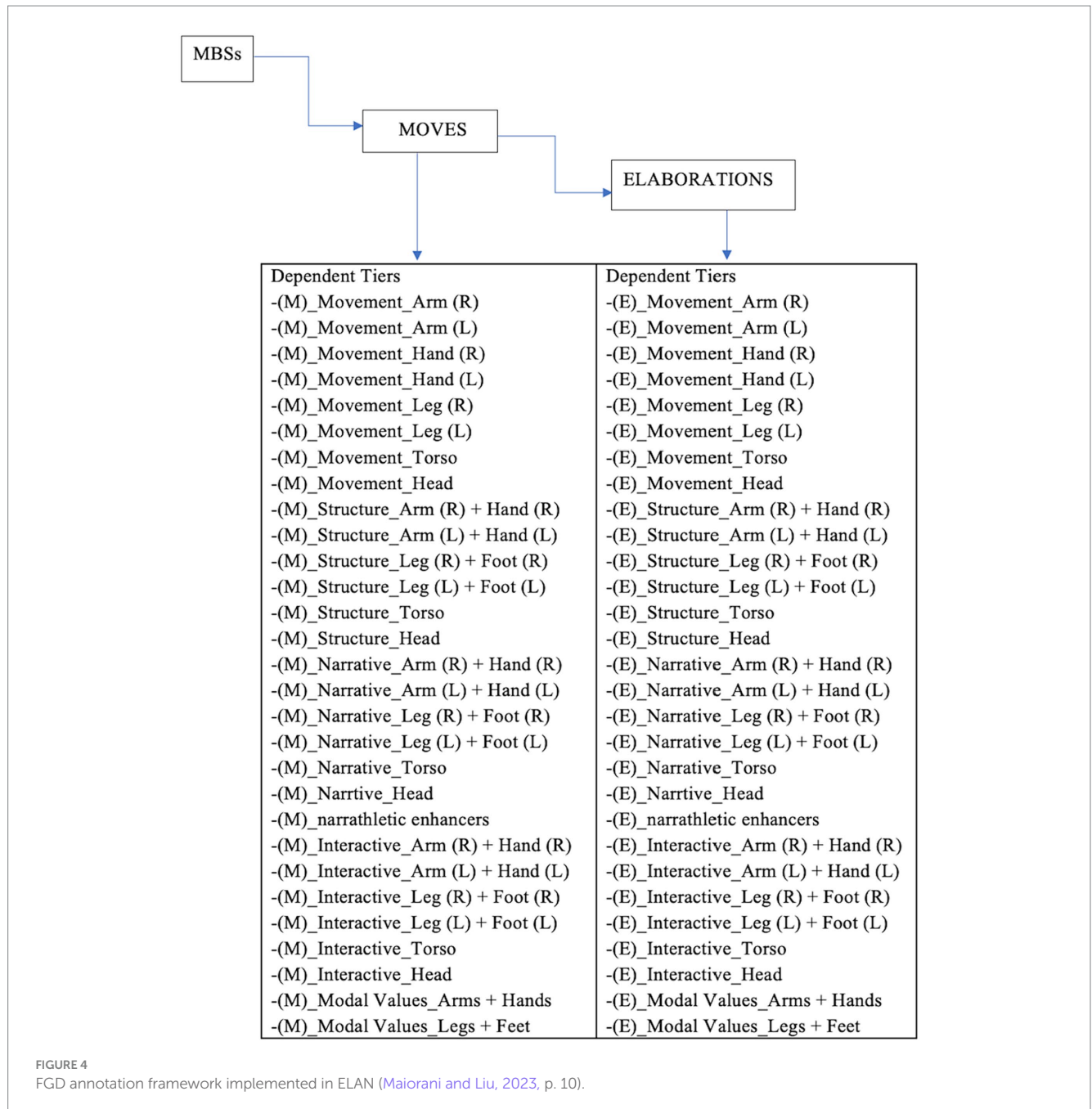




scope and capability. Bateman (2022, p. 42) highlights this problem with reference to work carried out before empirical data collection and analysis by stating that ‘many multimodal analyses were overly impressionistic, and that analyses tended in any case to be restricted to small-scale studies rarely capable of producing the degree of empirical robustness that would be necessary to improve on

impressionistic categories; even when the intuitions underlying such categories are generally sound, it is unlikely that they offer the last word on the precise treatments required’. The FGD application was therefore still restricted to a small case study, a first step that needed to be developed through more empirical work and larger, more complex data analysis.





The first development we achieved when we started working on the rich data provided by the live movement-capture sessions with the English National Ballet was the distinction between more local levels of annotation and more discursive ones. By implementing the FGD in ELAN for annotating our data, we found out that Moves and MBSs create different types of meaning at different levels that are then integrated through the movement flow: the same features that at the more local level of the Move have a specific structural function, at the level of MBS acquire a more discursive one. The analysis carried out on Moves through our annotation method reveals that at this level meaning is created more locally. The annotation of Moves provides three important sets of information that this minimal semantic unit delivers: two sets (starting and arrival) of narrative and interactive projections that the dancer

realises for the viewer to interpret within the context of the performance; the positions in which the dancer moves their articulators in relation to each other across space; the flow of relationships between *direction* and *orientation*, which not only determines the possible values to be attached to projections for the viewer (i.e., moving towards VS moving away from, going forwards VS going backwards, etc.) but also connects the more local meanings realised at the Move level to the syntactic choices observable at MBS level, where the direction of two consecutive Moves determines the type of MBS *trajectory*. These findings allowed us to gain an important insight into the mechanism of movement-based communication realised through choreography that the manual analysis simply based on the application of theory and tools did not allow us to uncover; the annotation of the rich data we collected

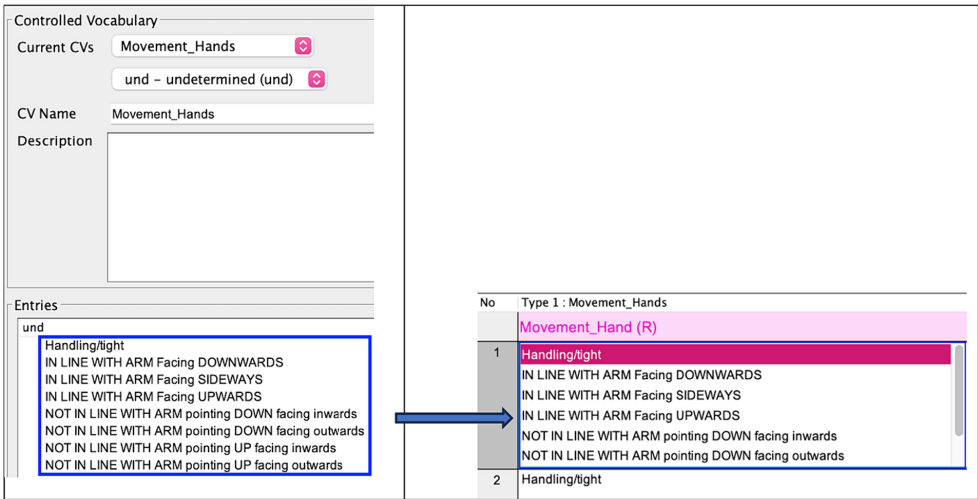


FIGURE 5  
Example of drop-down controlled vocabulary (Maiorani and Liu, 2023, p. 11).

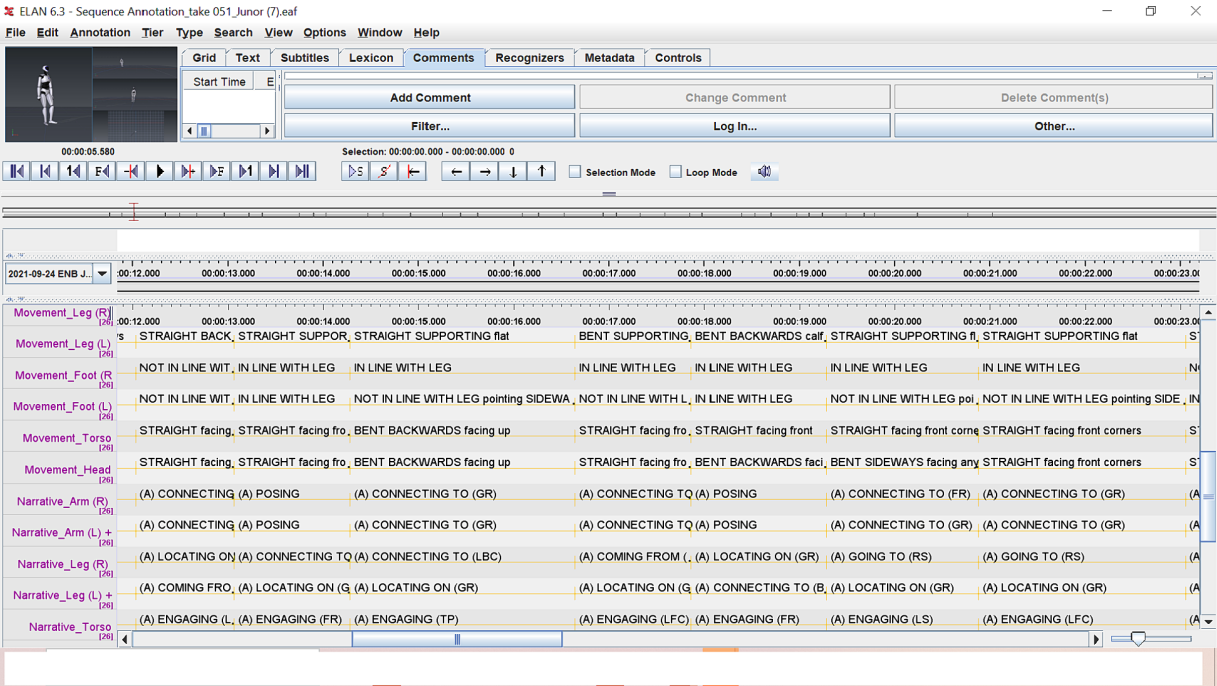


FIGURE 6  
Example of annotation with visualisation of a dancer's avatar.

through live movement-capture sessions put us in front of a multi-level discursive complexity that we would not have captured otherwise. The meanings expressed at the more local level of Move acquire a discursive flow in a more complex relation with the performance space at the level of MBS, which provides them with trajectories and highlights discursive patterns where bodies and space are integrated. Thus, whereas in previous manual analyses the focus had been predominantly on reading narrative and interactive projections on the basis of the theoretical framework underlying the FGD, thanks to empirical work our annotation had to take into

consideration the analysis of physical movement as a separate but integrated part of the analysis, foregrounding the importance of annotating the positions of the articulators with respect to each other and with respect to the physical space and to the dimensions of direction and orientation. The result was a systematic integration of the physical and the semantic data that are described in integration through the annotation within the same model and according to the same theoretical principles. The consideration of the materiality of dance, which involves also the integration of spatial features, direction, and orientation, led us to a discourse

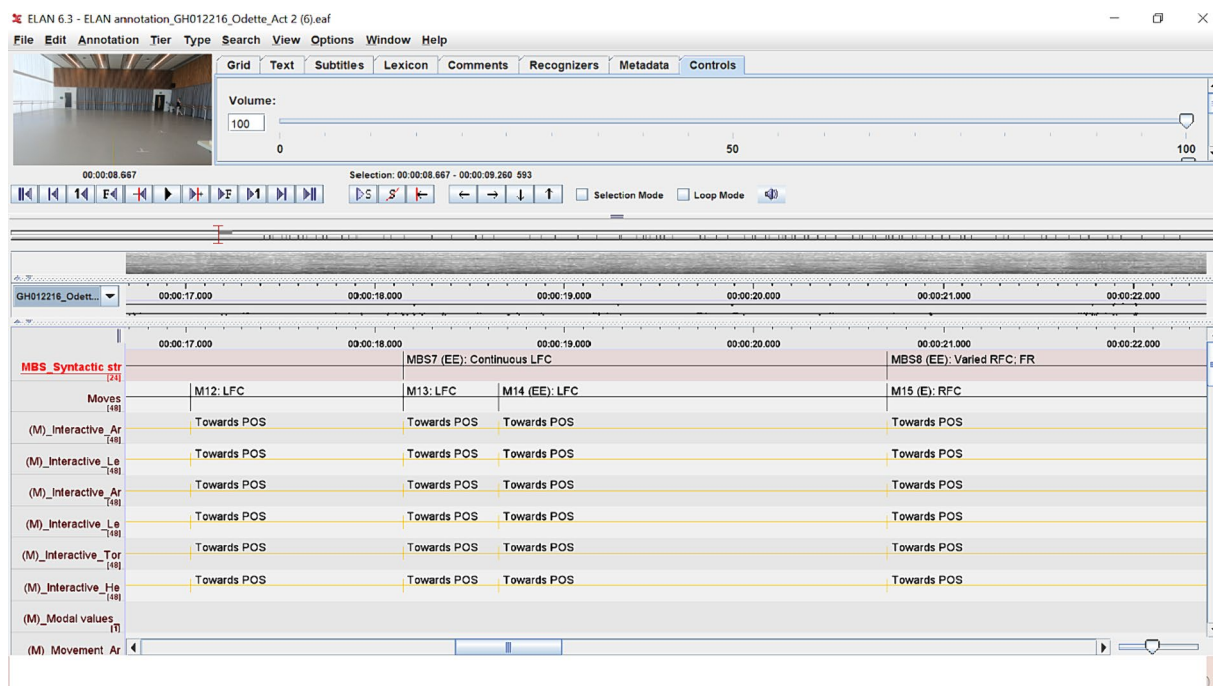


FIGURE 7  
Example of annotation showing the segmentation into Moves and MBSs.

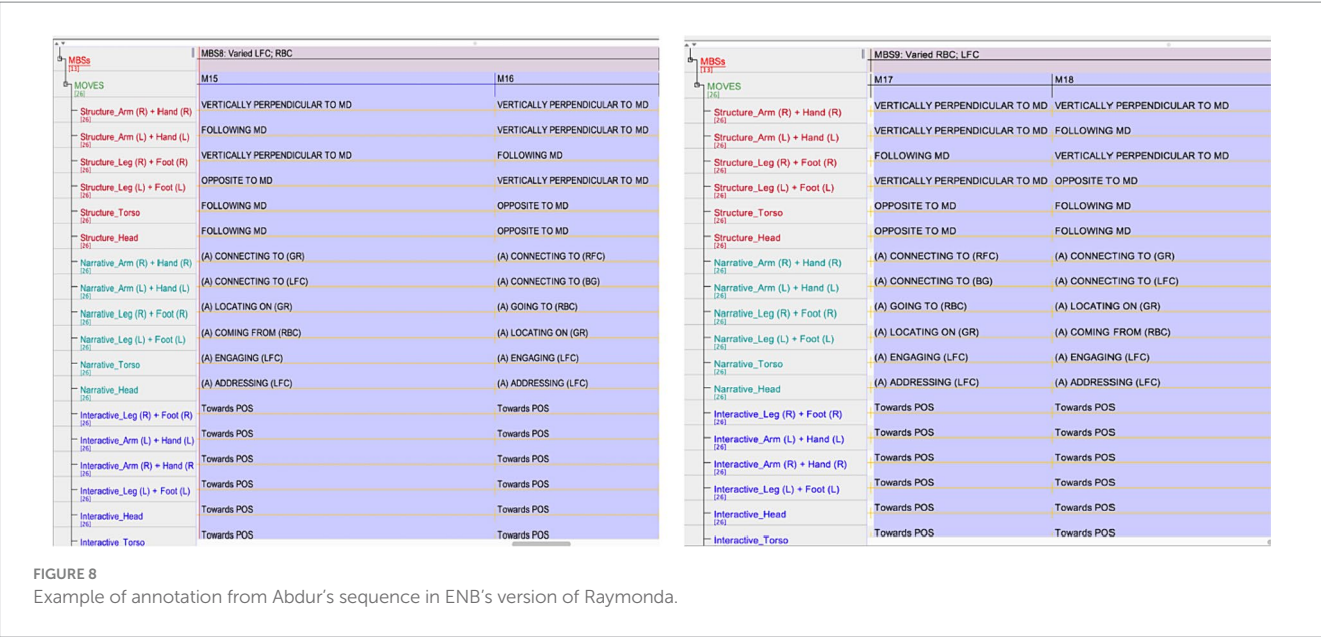
description that anchors the theoretically grounded visualisations of projections to the time and space of the physical dance performance phenomenon.

I will illustrate this point with a simple example of annotation taken from the English National Ballet's most recent production of *Raymonda*, a very traditional ballet from the classical repertoire first choreographed by Marius Petipa in 1987. Even in the most recent version, this traditional piece is based on an intricate and equally traditional love-triangle story against a romanticised historical setting, and it is therefore not too difficult for the audience to follow the flow of its scenes and the relationships amongst the characters based on the synopsis presented in the programme, which is supported by a very classical choreography. However, the extracts we annotated reserved us some surprises. Amongst the extracts we chose, there is one from a solo danced by Abdur, one of the three protagonists: in love with Raymonda, who is already engaged with his best friend, in the extract we analysed he finds himself alone with her and declares his feelings by dancing a solo variation. The annotation we carried out is exemplified in Figure 8, where a particular pattern is showing. The annotation includes four Moves and two MBSs and the figure shows annotations both of some physical movement structures and of some narrative and interactive projections. The more local annotation of the four Moves shows a considerable amount of repetition of specific physical movements that corresponds to an equally repetitive series of meanings: through narrative projections, the dancer interpreting Abdur forms repeated connections with Raymonda, who is sitting in the corner in the stage setup with which he is also repeatedly addressing and engaging. Interactive projections show that his interactions are entirely devoted to her and the stage space around her. However, the physical data annotation shows that the same types of narrative and interactive projections are being repeated alternatively in opposite directions and

maintaining the same orientation, thus indicating that Abdur is moving back and forth, towards and away from Raymonda, which impacts on the way narrative and interactive projections are perceived by the audience. The annotation itself offers a visualisation of this discursive pattern that develops across two MBSs, which we named 'mirrored pattern'. It also shows how the more local meanings at Move level are incorporated and shaped into a discursive strategy at the higher level of MBS. These patterns also made us realise that whereas movement *orientation* is important to capture the local value of narrative and interactive projections within Moves because it determines the perception of the narrative and interactive meanings realised by each set of movement structures, movement *direction* has a more discursive value because it incorporates those more locally determined meanings within a discursive flow that shows how those meanings can change in relation to the perception of the whole performance space surrounding the dancer. Whereas Abdur's unchanging Move orientation repetitively shows his focus towards Raymonda, his alternatively changing direction at MBS level shows the conflictual situation in which that focus is experienced by the character.

## Projecting conclusions

I have started this article with an overview of the way the notion of materiality is understood and used as a nexus for connecting the different components of complex phenomena approached by a variety of disciplines and research areas. The pattern that emerged from such an overview highlights how materiality is actually used, as foregrounded by Bateman (2022), as an external language for description that can be applied to several contexts where human experience manifests itself in and is carried out through multimodal



meaning-making socio-semiotic practices. The overview also showed that the notion of materiality helps anchoring theoretical advances to phenomenological studies, thus highlighting the importance of empirical data in any analytical activity across disciplines. The developments observed in other research areas were echoed by the developments evidenced in multimodal discourse analysis through an excursus of the work carried out within The Kinesemiotic Body project, which focused on movement-based discourse. In this specific case, the application of the Functional Grammar of Dance in the annotation and testing of empirical data led not only to the further development of this analytical model and its theoretical framework of reference, but also to a much better understanding of the complex structures that underlie movement-based discourse and their interaction with the contextual space in which communication happens, thus providing much stronger foundations for the extension of this type of analysis to forms of movement-based communication other than ballet and dance in general.

Focusing on materiality really means looking at the complexity of human experience and the processes through which it is shaped into semiotic constellations where configurations of modes work in interplay. Stage performances offer great examples of this complexity, involving music, dance, sung or recited text, movement, lighting, settings, costumes, all in need of more empirical investigation. When working on movement-based communication and in particular on movement-based performance, the 'materiality key' has opened the door to the integrated work of linguists, computer scientists, semioticians and engineers as it has provided a common ground for collecting, processing and analysing movement data under mutually understood and shared theoretical principles, and also for creating a common language for defining fundamental concepts. It has also highlighted the complex relationships occurring amongst the different factors that enable this type of communication where human bodies interact with space and its perception. Eventually, the project led to the creation of more effective ways of collecting, annotating and understanding movement data. Our work is still ongoing: one of the project's results was the creation of short videos where live-captured movement data is turned into avatars which

can be inserted in virtual stage set-ups that can be modelled *ad-hoc* for experiments on perception of how the body interacts with contextual space in communication. The avatars represent both female and male dancers and can now be visualised as carrying out projections when dancing across the virtual stage as the software is now capable of reading automatically the FGD annotations in ELAN. We can even select which types of projections to visualise and how to distinguish one type from the other. These visualisations are still undergoing some level of refinement but there is great potential for future applications in dance education for both dance students, professionals and general audience and for different forms of performance studies and movement-based communication analysis. The same principles of visualisation are currently being applied to the study of potential gender bias in the representation of avatars' movement in popular fighting games, thus extending the work started with The Kinesemiotic Body project to EDI issues related to the gaming world and relevant communities. These extensions of the work carried out by The Kinesemiotic Body project are possible precisely because the advances we made both in analysis and theory benefited from the focus on materiality as a descriptive language for unpacking and understanding the complexity of semiotic resources that work in interplay to produce dance discourse.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: <https://doi.org/10.17028/rd.lboro.c.6230502.v1>, Loughborough University Research Repository.

Ethics statement

The studies involving humans were approved by Loughborough University Ethics Review Sub-Committee. The studies were conducted in accordance with the local legislation and institutional requirements.



The participants provided their written informed consent to participate in this study.

## Author contributions

AM: Funding acquisition, Methodology, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. Funding received from Loughborough University through the UKRI.

## References

- Asher, N., and Lascarides, A. (2003). *Logics of conversation*. Cambridge: Cambridge University Press.
- Balmaceda, M., Högselius, P., Johnson, C., Pleines, H., Rogers, D., and Tynkkynen, V. P. (2019). Energy materiality: a conceptual review of multi-disciplinary approaches. *Energy Res. Soc. Sci.* 56:101220. doi: 10.1016/j.erss.2019.101220
- Bateman, J. A. (2019). Multimodality and materiality: the interplay of textuality and Textuality in the aesthetics of film. *Poetics Today* 40, 235–268. doi: 10.1215/03335372-7298536
- Bateman, J. A. (2022). Multimodality, where next? Some meta-methodological considerations. *Multimodal Soc* 2, 41–63. doi: 10.1177/26349795211073043
- Bateman, J. A., and Wildfeuer, J. (2014). A multimodal discourse theory of visual narrative. *J. Pragmat.* 74, 180–208. doi: 10.1016/j.pragma.2014.10.001
- Bateman, J., Wildfeuer, J., and Hiippala, T. (2017). *Multimodality: foundations, research and analysis. A problem-oriented introduction*. Berlin, Boston: De Gruyter.
- Björkqvall, A., and Karlsson, A. M. (2011). The materiality of discourses and the semiotics of materials: a social perspective on the meaning potentials of written texts and furniture. *Semiotica* 2011, 187–1/4. doi: 10.1515/semi.2011.068
- Bolens, G. (2022). Embodied cognition, Kinaesthetic knowledge, and Kinesic imagination in literature and visual arts. *Front Commun* 7:232. doi: 10.3389/fcomm.2022.926232
- Elyamany, N. (2023). A chronotopic approach to identity performance in musical numbers: a choreo-musical case study of ‘rewrite the stars’ and ‘this is me’. *Vis. Commun.* 22, 278–296. doi: 10.1177/1470357220974069
- Gherardi, S. (2017). “Sociomateriality in posthuman practice theory” in *The Nexus of practices: Connections, constellations, and practitioners*. eds. S. Hui, E. Shove and T. Schatzki (London-New York: Routledge), 38–51.
- Gibbons, A. (2012). *Multimodality, cognition, and experimental literature*. London: Routledge.
- Kell, C. (2015). “Making people happen”: materiality and movement in meaning-making trajectories. *Soc. Semiot.* 25, 423–445. doi: 10.1080/10350330.2015.1060666
- Kress, G. (2010). *Multimodality: A social semiotic approach to contemporary communication*. London-New York: Routledge.
- Lacković, N., and Popova, B. (2021). Multimodality and sociomateriality of lectures in global universities’ media: accounting for bodies and things. *Learn. Media Technol.* 46, 531–549. doi: 10.1080/17439884.2021.1928694
- Lee, T.-K. (2014). Translation, materiality, intersemioticity: excursions in experimental literature. *Semiotica* 202, 345–364. doi: 10.1515/sem-2014-0044
- Leonardi, P. M., and Barley, S. R. (2010). What’s under construction Here? Social action, materiality, and power in constructivist studies of technology and organizing. *Acad. Manag. Ann.* 4, 1–51. doi: 10.5465/19416521003654160
- Maiorani, A. (2017). “Making meaning through movement: a functional grammar of dance movement” in *Mapping multimodal performance studies*. eds. M. G. Sindoni, J. Wildfeuer and K. L. O’Halloran (London/New York: Routledge), 39–60.
- Maiorani, A. (2020). Selling the past and the present alike: streaming ballet for live audiences during lockdown. *J Int Cult Arts* 1, 1–9. doi: 10.46506/jica.2020.1.2.001
- Maiorani, A. (2021). *Kinesemiotics: Modelling how choreographed movement means in space*. London-New York: Routledge.
- Maiorani, A., Bateman, J. A., Liu, C., Markhabayeva, D., Lock, R., and Zecca, M. (2022). Towards semiotically driven empirical studies of ballet as a communicative form. *Human Soc Sci Commun* 9:429. doi: 10.1057/s41599-022-01399-8
- Maiorani, A., and Liu, C. (2023). The functional grammar of dance applied to ELAN annotation: meaning beyond the naked eye. *J World Lang.* doi: 10.1515/jwl-2023-0050
- Maiorani, A., and Wegener, R. (2022). “Challenging instantiation in modelling movement-based multimodal communication” in *Empirical evidences and theoretical assumptions in functional linguistics*. eds. E. Asp and M. Aldridge (London – New York: Routledge), 151–169.
- Meissl, K., Sambre, P., and Feytaerts, K. (2022). Mapping musical dynamics in space. A qualitative analysis of conductors’ movements in orchestra rehearsals. *Front Commun* 7:733. doi: 10.3389/fcomm.2022.986733
- Merriman, P., and Jones, R. (2017). Nations, materialities and affects. *Prog. Hum. Geogr.* 41, 600–617. doi: 10.1177/0309132516649453
- Mouard Ruiz, E. (2021). En movimiento: audiodescripción del acuario del arrecife de coral. Universidad de Granada. Departamento de Traducción e Interpretación. Available at: <http://hdl.handle.net/10481/69420>
- Pantidos, P., Valakas, K., Vitoratos, E., and Ravanis, K. (2010). The materiality of narrative spaces: a theatre semiotics perspective into the teaching of physics. *Semiotica* 2010, 305–325. doi: 10.1515/semi.2010.062
- Petrilli, S. (2008). Bodies and signs: for a typology of semiotic materiality. *Am J Semiotics* 24, 137–158. doi: 10.5840/ajs200824422
- Prové, V. (2022). Measuring embodied conceptualizations of pitch in singing performances: insights from an OpenPose study. *Front Commun* 7:987. doi: 10.3389/fcomm.2022.957987
- Putnam, L. (2014). Unpacking the dialectic: alternative views on the discourse-materiality relationship. *J. Manag. Stud.* 52:115. doi: 10.1111/joms.12115
- Schatzki, T. (2010). Materiality and social life. *Nat Cult* 5, 123–149. doi: 10.3167/nc.2010.050202
- Sindoni, M. G. (2022). Traiettorie della multimodalità: gli snodi teorici e i modelli applicativi. *Ital LinguaDue* 14:3597. doi: 10.54103/2037-3597/2
- Tyrer, C. (2021). The voice, text, and the visual as semiotic companions: an analysis of the materiality and meaning potential of multimodal screen feedback. *Educ. Inf. Technol.* 26, 4241–4260. doi: 10.1007/s10639-021-10455-w
- Vidal Caramonte, M. Á. (2022). *Translation and contemporary art: transdisciplinary encounters*. New York: Routledge.
- Wilf, E. (2011). Sincerity versus self-expression: modern creative agency and the materiality of semiotic forms. *Cult. Anthropol.* 26, 462–484. doi: 10.1111/j.1548-1360.2011.01107.x
- Wu, X. (2022). *Space and practice: a multifaceted understanding of the designs and uses of “active learning classrooms”*. UNSW: Sydney.

## Software

ELAN (Version 6.2) [Computer software]. (2021). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Available at: <https://archive.mpi.nl/tla/elan>.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.





## OPEN ACCESS

EDITED BY  
Claudia Lehmann,  
University of Potsdam, Germany

REVIEWED BY  
Karl-Heinrich Schmidt,  
University of Wuppertal, Germany  
Assimakis Tseronis,  
Örebro University, Sweden

\*CORRESPONDENCE  
Dayana Markhabayeva  
✉ dayana1@uni-bremen.de

RECEIVED 01 December 2023  
ACCEPTED 08 March 2024  
PUBLISHED 28 March 2024

CITATION  
Markhabayeva D and Tseng C-I (2024)  
Multimodal cohesion and viewers'  
comprehension of scene transitions in film: an  
empirical investigation.  
*Front. Commun.* 9:1347788.  
doi: 10.3389/fcomm.2024.1347788

COPYRIGHT  
© 2024 Markhabayeva and Tseng. This is an  
open-access article distributed under the  
terms of the [Creative Commons Attribution  
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic practice.  
No use, distribution or reproduction is  
permitted which does not comply with these  
terms.

# Multimodal cohesion and viewers' comprehension of scene transitions in film: an empirical investigation

Dayana Markhabayeva<sup>1\*</sup> and Chiao-I Tseng<sup>2</sup>

<sup>1</sup>Faculty of Linguistics and Literary Studies, University of Bremen, Bremen, Germany, <sup>2</sup>Department of Applied Information Technology, University of Gothenburg, Gothenburg, Sweden

This paper presents three empirical studies that unravel how the devices of multimodal cohesion support viewers' narrative interpretation of scene transitions in film. The linguistics-informed method of cohesion analysis in film uncovers the establishment of cohesive ties between characters, objects, settings and characters' actions. Previous studies using eye-tracking and comprehension tests already indicate the significance of multimodal cohesion in people's comprehension of background settings within a continuous scene. The present paper investigates further whether film cohesion impacts viewers' story comprehension across different scenes and settings. Moreover, it also explores whether the spatio-temporal relations between scenes is a significant factor, along with cohesive devices, in viewers' scene comprehension. Methodologically, we create contrasting film situations by manipulating cohesion structures and spatio-temporal orders of scenes. Our comparative analyses of viewers' comprehension of these different film situations reveal that the presence of cohesive cues significantly can influence viewers' accurate scene comprehension. Through testing the inter-relation of cohesion, spatio-temporal order, characters' intention and viewers' time perception, this paper offers new avenues for further exploration of space, time and coherence in film.

## KEYWORDS

film, cohesion, scene transition, multimodal discourse analysis, narrative comprehension, spatio-temporal relations

## 1 Introduction

In the history of film studies, the ways of how the viewers are carried from shot to shot and scene to scene as well as the effects of different types of scene transitions have been frequently investigated. Eisenstein (1969) explored how filmmakers' combinations of different shots and scenes can lead the viewers to interpret meaning in particular ways. In the 1970s, transitions of shots and scenes were systematically analyzed by the semiotician Christian Metz. Metz (1974) *Grande Syntagmatique* proposes eight types of cinematic syntax, namely, the transitions between film shots and scenes. Following the pursuit of Metz to create a generalized modeling of scene and shot transitions, the recent study of Bateman (2007) and Bateman and Schmidt (2012) proposed a *Grande Paradigmatique*, which maps out a more comprehensive set of semantic relations between different shots such as different types of spatio-temporal and logical relations.

While the semiotic theories by Bateman focus on shot-based semantic relations, other scholars have explored another type of mechanism, namely, *cohesion*, addressing how verbal, visual and audio elements *within* film shots are tied together to signal the coherent flow of film narratives across scene changes.

Cohesion is originally a linguistic concept. It refers to a set of semantic relations in text which enable the interpretation of meaning coherence. In text linguistics (Halliday and Hasan, 1976), text as a coherent whole is the result of cohesive devices at work. Halliday and Hasan (1976, p.12) posits that “cohesion is a relational concept; it is not the presence of a particular class of item that is cohesive, but the relation between one item and another”.

In the context of cinema, the film theorist Bordwell (2008) provides an exploratory account of how different types of audiovisual cohesion function to carry viewers across scene transition and how patterns of film cohesion unravel viewers cognitive comprehension activities. Bordwell exemplifies, for instance, how cohesive relation is established when the re-occurrence of a same object or characters in two different scenes cues the viewer to interpret the coherent narrative flow. The recent empirical studies of scene comprehension (Loschky et al., 2015b) also investigate the elements within and across shots that lead to coherent scene perception. This paper, employing Bordwell’s definition of *scene* and insights from cognitive studies, examines scene transition as a disruption in space and time, namely, a change of event location and the break of continuous events.

Systematically employing the linguistic concept of textual cohesion, the two works of Tseng (2012, 2013) extend Bordwell’s attempt to identify internal cohesive structure and propose a more systematic framework of film cohesion. In linguistic analysis, the analytical tools of cohesion are used to describe the “repetition” and “re-occurrences” of linguistic patterns, with which a text holds itself together as a unit of communication. Along the same lines, the multimodal cohesion analysis unravels how the “repetition” and “re-occurrences” of narrative elements such as people, places, objects and actions, whether identified in the visualtrack (e.g., visible figures or as written names on the screen) or in the audiotrack (e.g., spoken names or sounds and music that represent certain identities), are cued to the viewers for interpreting the narrative coherence within and across shots.

While the framework of multimodal cohesion has been applied to other media such as TV series, comics, other graphic novels (Tseng and Bateman, 2018; Tseng et al., 2018; Drummond and Wildfeuer, 2020) and interactive narratives (Tseng and Thiele, 2022), there has not been sufficient empirical investigations of just which verbal, visual and audiovisual cohesive cues in film play the dominant, pivotal role of facilitating the seamless connection between the storytelling units.

One of the first empirical attempts for triangulating the multimodal cohesion framework and the viewer’s cognition and memory is conducted by Tseng et al. (2021). The authors use comprehension tests and eye-tracking experiments to compare viewers’ attention and narrative interpretation of film sequences, either with or without cohesive cues crucial for the viewers’ comprehension of the specific settings within a scene. They use film sequences extracted from *The Birds* and a *Monty Python* sketch for the experiments and their findings indicate the significant role of cohesive devices for the viewers’ narrative comprehension and gaze-behavior. The findings open up more questions as to whether cohesion still plays a role in the more complex scene transitions, whether cohesion in film influences the way the viewers interpret the continuity of spatio-temporal and logical relations as those

theorized by Bateman (2007) and whether cohesion and spatio-temporal relations are related to viewers’ perception of intention and time. This paper will precisely extend the previous empirical endeavor to address these open questions.

The paper is structured as follows: Section 2 exemplifies the analysis of multimodal cohesion and how the cohesive structures, termed *cohesive chains*, reflect the viewers understanding of the presentation and re-occurrences of characters, objects and settings. Drawing on the multimodal cohesion analysis, Section 3 presents the empirical studies we conduct to triangulate the multimodal linguistic framework with the viewer’s cognitive process. Several cognitive studies have endeavored to address how the audiences’ coherent narrative comprehension is steered by film narrative and technical features such as continuity editing (Smith, 2012), event recognition (Zacks, 2015) and scene construction (Loschky et al., 2015a). Our studies of multimodal cohesion complement the previous cognitive studies through providing a semiotically formulated model of interpretation. As we will see in the next section, this semiotic-textual level of analysis offers a more fine-grained yet systematic investigation interconnecting the functions of film technical features, narrative elements, semantic structures and the overall contextual coherence.

## 2 Analysing multimodal cohesion in film

The framework of multimodal cohesion (Tseng, 2013) provides a powerful discourse semantics for examining cohesive ties between film elements within and across shots and scenes. It was formulated drawing on the discourse semantic model of identification, which was developed for the analysis of natural language (Martin, 1992). In the linguistic analysis, the choices of the identification system realize the identity presentation and re-occurrence of people, places and things throughout a text. The structures of identification, namely, how relevant people, places and things are actually tracked, then highlight the textually constructed unity of any particular text. Tseng (2013) applied the discourse semantic framework to film. In this way, the framework captures not only the area of semiotic work shared across language and film but also the differentiation of the filmic cohesion analysis from the linguistic analysis.

The multimodal cohesion system developed for film is shown in Figure 1 represented as a system network. System networks are used in systemic functional linguistics to show the abstract paradigmatic “choices” available for language users drawn from the meaning potential of their language (Halliday and Matthiessen, 2013). In the film system, the network in Figure 1 shows the functional mechanisms for cuing identities of characters, objects and settings as a film unfolds. In the system network, contrasting options are collected together into individual systems of choice: for instance, in the system of [presenting/presuming], only one of the two features may be selected at a time. Certain feature selections then also lead on to finer classifications. For example, in the case of the choice [gradual], the system leads on to a further dependent, i.e., finer, choice between [dynamic] and [static]. It is also possible for several dimensions of classification to be pursued in *parallel*: such systems are called simultaneous systems and are grouped with a curly right-facing bracket. In Figure 1, for example, choices need to be made

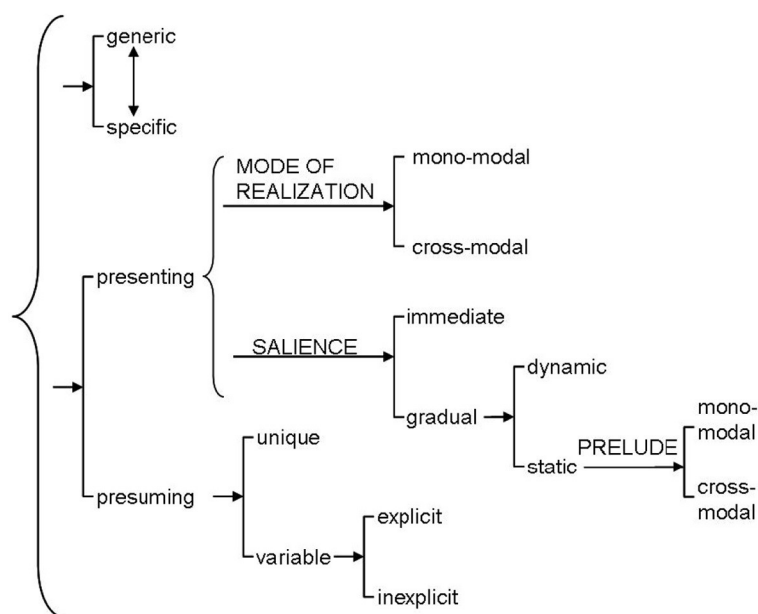


FIGURE 1

The system of multimodal cohesion in film developed by Tseng (2013).

from the features presented by *both* the systems [generic/specific] and [presenting/presuming] for a complete description.

We exemplify the process of constructing a multimodal cohesive analysis of film using a scene of Nolan's (2000) *Memento*. The film is a thriller, depicting the main character, Leonard Shelby, an insurance investigator, suffers from short-term memory loss and uses notes and tattoos to hunt for the man he thinks killed his wife, which is the last thing he remembers. The segment we exemplify here is a scene when Leonard goes into a tattoo shop. We also employ this segment below in one of our experimental studies.

Figure 2 shows the scene transition from street view to the indoor setting of a tattoo shop. It includes selected shots that can best depict the location transition across the outdoor and indoor scenes. Shot 1a shows the front door of a shop. A small orange sign at the bottom left shows it is a tattoo shop. Within the same shot, a car is seen and heard squeaking and stopping abruptly in front of the shop (shot 1b). It is then cut to shot 2, the closeup of Leonard, who is seen looking at a white object. The point-of-view shot in shot 3 then shows the note he is reading, with *Tattoo Fact 6: car license* written on it. A closeup in shot 4 shows the shop's name, *Emma's TATTOO*. It is followed by the transition into the indoor scene of the shop. Shot 5 shows the closeup of the tattooist's hands tattooing the same written text of the note seen in shot 3 on someone's skin. Shot 6 and shot 7 reveal that Leonard is the one being tattooed. Throughout shots 5 to 7, the audience can hear the continuous tattooing sounds. The second character, Teddy, entered the room in shot 8, greeting Leonard: "Lenny!". While in shot 9, Leonard lifts his head, seeming not remembering who the person is, in shot 10, the tattooist yells at Teddy: "It is private back here. Wait out there". In shot 11, Teddy looks frustrated but goes out to wait in another room in the same tattoo shop, shown in shot 12. The

same tattoo setting is suggested by the background tattoo images and symbols on the walls of the room. In shot 13, Leonard and the tattooist both came out. Shots 14 and 15 construct shot/reverse shot showing the conversation between the two characters in the room.

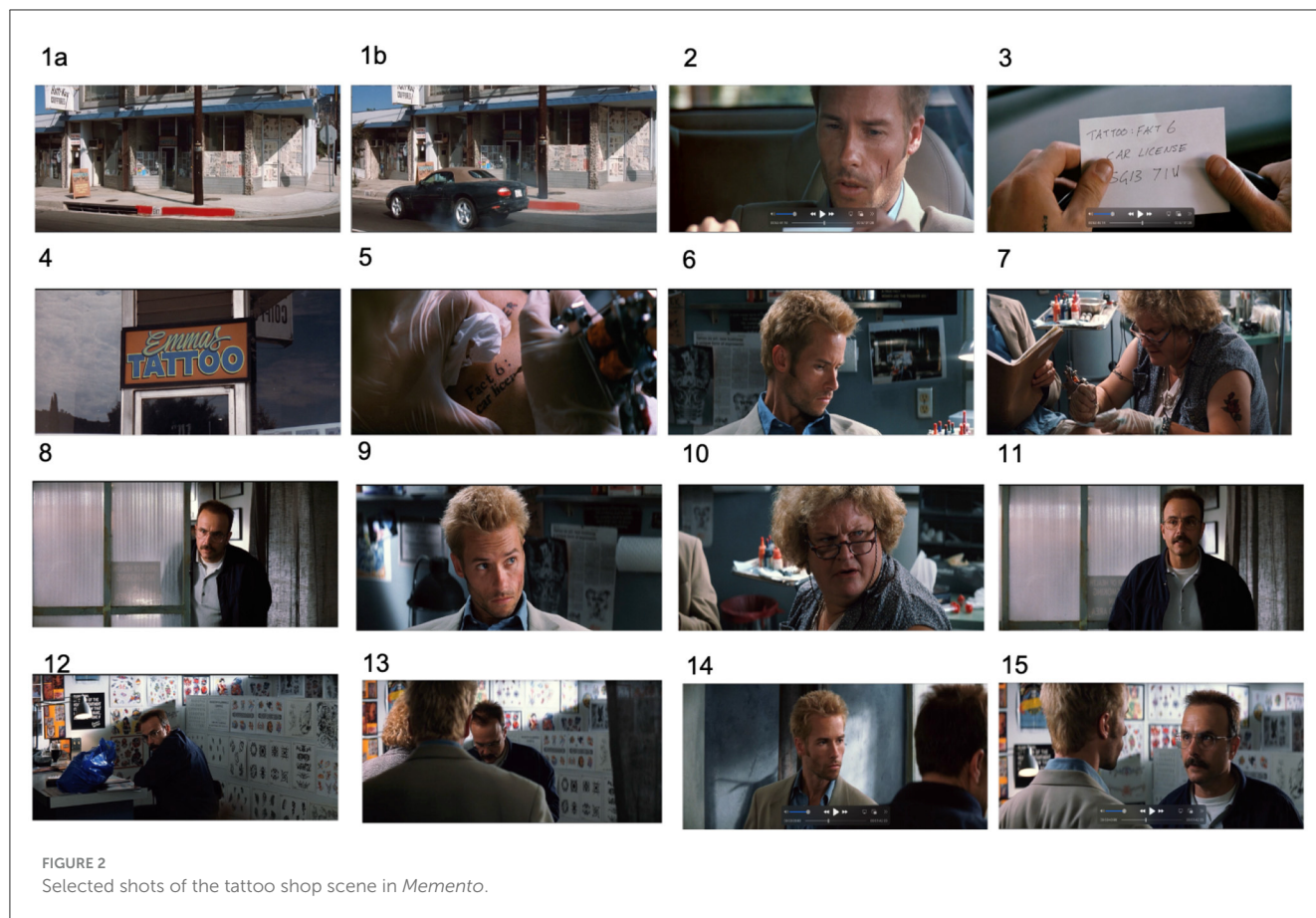
Focusing for the purposes of illustration on the setting of *Tattoo shop*, we can describe the cohesive devices for presenting and tracking the tattoo shop based on the instantiation of features from the system network of Figure 1.

In shot 1, the front door of a shop is seen from a street view. For the viewers who notice the *Tattoo* sign written on the orange board at the left corner, the specific identity of the tattoo shop is immediately established. This is therefore a case of [presenting] rather than [presuming]. As the shop is specified as *Tattoo shop* right at the outset, the cohesive devices at work are therefore [specific] from the continuum [generic - specific] and [immediate] salience.

In shot 3, the written text *tattoo* on the paper held in Leonard's hands is the multimodal re-occurrence of the tattoo shop. Although *tattoo* does not directly refer to the shop setting, it is a *hyponym* of tattoo shop, cohesively related to the previously seen tattoo setting. Hence, this is the case of [presuming] that track the same identity of tattoo shop.

Similarly, in shot 4, the front door of the shop with the shop name *Emma's Tattoo* cohesively cues the viewers back to the tattoo note. Here the cohesive devices [presuming] and [explicit] reappearance are at work to track of the tattoo shop.

From shot 5 onwards, re-occurrences of the theme *Tattoo* and *Tattoo shop* are visualized through more sets of multimodal elements: the female tattooist's tattooing Leonard's thigh, the continuing tattooing sounds and the background tattoo pictures in the room from shots 12 to 15.



In other words, all these verbal, visual and audio cues are cohesively tied together to signal the concept of tattoo/tattoo place; and each re-occurrence of a cohesive element is related to preceding occurrences by specifically labeled cohesive ties showing the tracking strategy involved.

Whereas cohesive ties relate pairs of cohesive elements, sequences of element re-occurrences and the classified cohesive ties between those occurrences are structured into *cohesive chains*, which show textual development of narratively significant characters, objects and settings across larger portions of film sequences. The overall cohesive chains structured from the tattoo shop scene is displayed in Figure 3. Here we can see that the tattoo shop chain interlinks the multimodal realization of the elements we discussed above. This chain starts with *Tattoo* in written text and the visual figure of the tattoo shop, annotated as [V]. It is then linked by an upward pointing arrow from both written text *Tattoo* in shot 3. The arrows refer to the semantic relation of anaphora, which ties the pairs of cohesive narrative elements together. Along the same lines, the continuing multimodal cohesive chain links together the written text *Emma's Tattoo*, audio (tattooing sounds) and visual (indoor shop) elements of the setting *tattoo shop*.

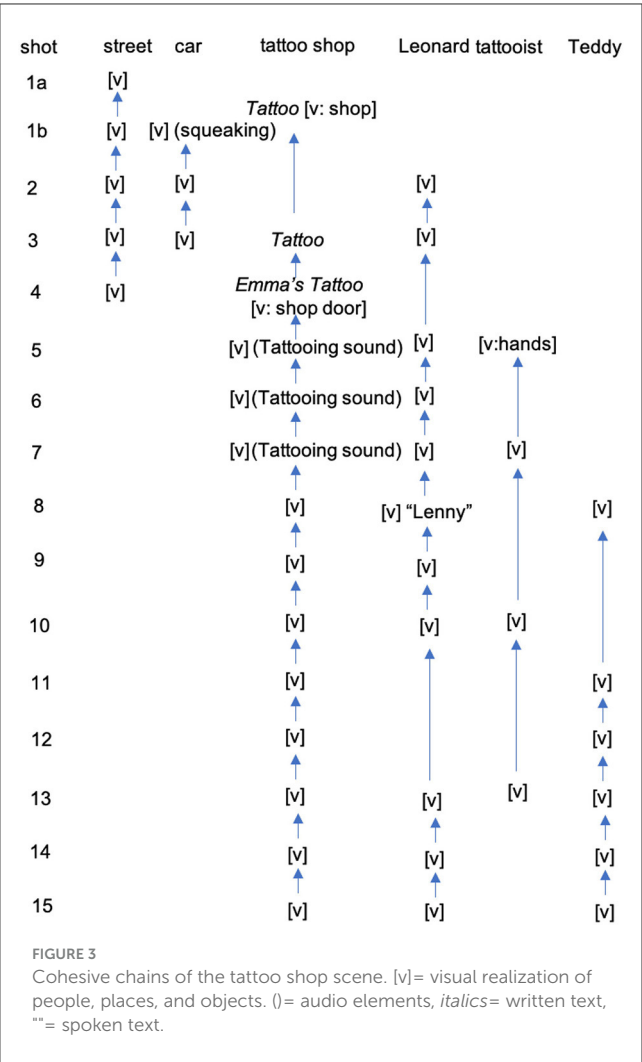
Similar to the cohesive chain of *tattoo shop*, the chain of Leonard shows how this character is visually presented (in visual image [v] in shot 2) but reoccurred multimodally in the following shot—in shot 8, his identity is realized in spoken text when Teddy called his name. Moreover, the cohesive chain of *car* also shows a multimodal presentation of the object (shot 1b) – it is not

only seen but also heard when the car breaks with a squeaking sound. Hence, the first element of the *car* chain is annotated as [v](squeaking).

Moreover, in research work on verbal texts (Hasan, 1984) as well as film (Tseng, 2008), it has been observed that such chains and, in particular chain *interactions*, appear to be more revealing of a text's organization than elements that occur in relative isolation. Interactions between chains occur whenever elements of distinct chains are brought together within the depiction of a single action or event. Thus, although any element in a textual artifact typically enters into a large number of cohesive links with other elements, it is the elements participating in chain interactions that are constructed as being textually “significant”. This constructs a useful method for selecting from all the cohesive ties potentially available in a text just those collections of ties that are hypothesized to be most likely to play a role in guiding the viewers' narrative interpretation. That is, a viewer does not need to attend to “everything” that is audio-visually on offer, but rather will be guided to attend to those elements that contribute to interacting chains. For example, before the scene transition across shot 4 and 5, namely, from the outdoor to the indoor scenes, the cohesive chains of *street* and *car combine/interact* with the chains of character Leonard and setting *tattoo shop* to construct a coherent event that might be glossed in natural language as follows:

On a *street* in front of a *tattoo shop*, a *car* driven by *Leonard* stops in front of the shop door.





After the scene transition into the indoor setting, the cohesive chains allow the construction of sequences of three further events:

While Leonard in the *tattoo shop* is being tattooed by a *tattooist*, *Teddy* comes in and interacts with him.

We predict that viewers are likely to see such generalized events across the scene transition based on the audiovisual material they engage with. Therefore, the deployment of the material possibilities of film itself serves a central role in guiding a film's reception. Hence, we predict that the *interaction* of cohesive chains (Hasan, 1984), namely, how elements in cohesive chains are combined/co-occur in each shot, should lead to the interpretation path of the generalized events across scene transition.

We explain how we experimentally investigated this in the following questions.

### 3 Toward experimental investigation

We have predicted that multimodal cohesion analysis can reveal how the filmic elements presented and maintained in a film

sequence lead the viewers to interpret particular 'events' across scene transitions on the basis of the available cohesive cues.

In the remainder of this paper, we investigate the empirical support for such a close association of cohesive patterning and narrative interpretation. In this pursuit, we employ the methodology of selecting film sequences and systematically modifying those sequences so that different patterns of cohesion are established. The two sequences, original and manipulated, are then shown to different groups of viewers. We then measure and compare the comprehension and engagement by the two groups of participants.

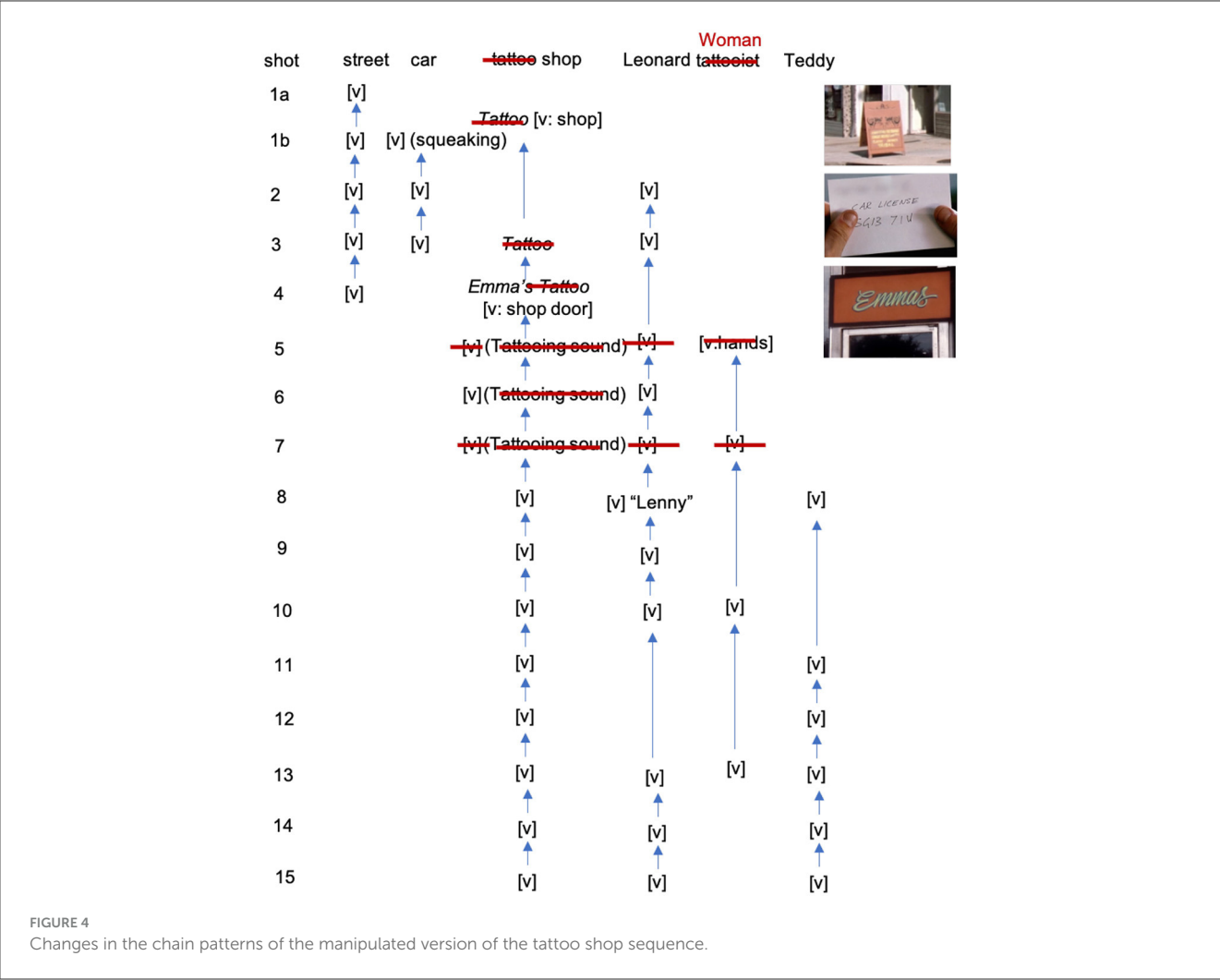
The measurement was conducted by providing participants with questionnaires designed to evaluate their understanding of the observed events. Three studies testing the functions of cohesive cues in events across scene transitions were performed. The first uses the sequence of the tattoo shop scene in *Memento* analyzed above. The second study employs the same method but factors in the aspect of spatio-temporal cues to test the viewers scene transition in the beginning sequence of *Memento*.

While the results of the second study revealed little effect for the spatial-temporal order, we speculate that this is because it is a puzzle film which begins with loosely connected scenes. This kind of challenging patterns of scene transitions in the film beginning is typical of puzzle film genres (Bateman and Tseng, 2013). This motivated us to conduct a third study using a different film with a structure distinct from the complex, puzzle structure of *Memento*. We chose Ephron's (2009) *Julie & Julia*. The movie has simple, linear structure. It portrays the lives of two women, Julie Powell and Julia Child. Julie finds herself in a career rut and decides to challenge herself by embarking on a journey to cook all the recipes from Julia Child's cookbook. She documents her experiences in a blog and discovers a new passion for cooking. The segment we selected for the experiment is a scene in which Julie goes to a butcher's shop to buy ingredients after her failed attempt to cook a dish from Julia's book. In the third study, in addition to selecting a film with a different structure, we also expanded the scope of the measures. We used more fine-grained scales for comparing degrees of correctness of participants' responses and included the measures of participants' confidence level for their responses. We also tested viewers' interpretation of the main character's intention and the length of event time.

In other words, the presentation of the three studies demonstrates the process of our step by step investigation into the complex configuration of cohesive and structural factors in the viewers' narrative comprehension of events and scene comprehension.

### 3.1 Study 1: "tattoo shop" scene in *Memento*

For the manipulation of the sequence analyzed above, we focused on the scene transition from the street view to the specific tattoo shop. As suggested in Figure 3, the original sequence encompasses sufficient multimodal cohesive cues for explicitly identifying just what kind of shop Leonard is in after the scene transition. For testing the functions of these



multimodal cohesive cues, we subtly removed the specific cues that indicate the identity of *tattoo* shop. That is, we blurred the written signs (on the orange board in shot 1, on the paper in shot 3 and on shop door in shot 4) identifying it as a tattoo shop. We replaced the tattooing sound with some generic background music and we cut out shot 5 and shot 7, the closeups of the tattooing actions. As the two shots removed for the experimental version are close-ups of the tattooing actions, the manipulation does not lead to the loss of other significant story information.

Except for the texts, sounds and actions about the tattoo, the manipulated sequence is identical to the original sequence. Our hypothesis is that removing the cohesive connections in this way should nevertheless disrupt the viewers' narrative comprehension of scene transition from street view to the specific indoor location.

Figure 4 shows the audiovisual cohesive chain analysis of the modified sequence. The removal of the tattoo text in shots 1, 3 and 4 results in the change of the setting from a specific named tattoo shop to a generic indoor space. It could still be recognized as a shop due to some visual elements such as the

orange board in shot 1, *Emmas* in shot 4, which are usually recognized as signs for a shop. In terms of multimodal cohesion and the classification system of Figure 1, therefore, the modification undertaken at the discourse level was a change in presentation strategy for the shop scene from a [specific] to [generic] shop. The manipulation also resulted in the change of the tattooist chain - as we cut away shot 5 and shot 7 when the woman is seen specifically as a tattooist, the woman shown in shots 10 and 13 then changes to a [generic] woman in the indoor space.

### 3.1.1 Hypothesis

To assess whether the manipulations indeed disrupted participants' comprehension of the excerpt, we tested the specific hypothesis:

- Viewers of the manipulated versions will be less certain about the specific identities of the shop, even though the relevant visual elements inside the shop (i.e. tattoo pictures on the wall) are still readily accessible on screen.

TABLE 1

	Cued	Uncued	Total
Comprehended	23	13	36
Not comprehended	0	9	9
Total	23	22	45

*Memento* tattoo shop study: number of participants with correct or incorrect answer to the question in group 1 (cued version) and group 2 (uncued version).

### 3.1.2 Experiments

This hypothesis was investigated by having participants answer the following questions immediately following their viewing of the *Tattoo shop* segments:

- “Where is the setting of the indoor place”?

The comprehension test was conducted at the University of Bremen, and participants ( $n = 45$ ) were undergraduate students who had not seen the film before the experiment. The participants were divided into two groups. Group 1 ( $n = 23$ ) watched the original versions (i.e., the cohesively “cued” versions) of the two sequences, while Group 2 ( $n = 22$ , “uncued”) viewed the manipulated versions with cohesive cues removed. Fisher’s exact test was used to evaluate statistical significance of dependencies between the cohesion status (cued *vs.* uncued) and viewers’ interpretations of the location (correct *vs.* incorrect), with  $p < 0.05$  considered significant.

### 3.1.3 Results

Table 1 presents the test results. All 23 participants in Group 1 who watched the cued version were aware of the specific identity of the tattoo shop, while only 13 participants from Group 2, who watched the uncued version (without cohesive cues), answered the question correctly. The 6 participants who were not certain about the location gave answers varying from a generic room to an office. Fisher’s exact test shows a significant association between the independent variable “cohesion” (cued/uncued) and the dependent variable “establishment of the setting’s identity” (correct/incorrect) ( $p = 0.0006$ ). Thus, although it is certainly the case that viewers of the uncued version might be able to guess the kind of shop involved correctly based on the pictures of tattoo patterns in the background (in shots 12–15), the question interrogated here is whether the manipulation makes a difference. The results demonstrate that the cued and uncued versions indeed differ significantly in comprehension.

While the previous empirical study of multimodal cohesion by Tseng et al. (2021) focuses on setting interpretation within one continuous scene, our result above re-endorses the empirical ground of cohesive setting across a scene change. The test design was further expanded in the second study to include the factor of spatio-temporal cues between scenes.

## 3.2 Study 2: the beginning four scenes of *Memento*

In order to show how the deployment of cohesive and spatio-temporal relation can interact and guide narrative interpretation, in the second study, we applied the same method of multimodal cohesive analysis to the beginning sequence of *Memento*. It is the first seven minutes of the film composed of a two-track alternating sequence of four scenes. The detailed cohesion analysis of the four scenes are provided by Tseng (2013). Here we focused on the comprehension tests of the transition of the four scenes, which we simply label S1, S2, S3 and S4, respectively, in order to emphasize their location and inter-relations in the film. Figure 5 shows the transition of the four scenes and the shots before and after the transition points. The changes of these four scenes are very clear for viewers in that their boundaries are signaled through fade-outs and fade-ins, which give the viewer explicit cues for recognizing that a new narrative segment may be beginning.

The first scene, S1, is presented in color and runs behind the opening credits. It depicts events in which Leonard shoots Teddy dead. This scene runs in reverse: i.e. the film is actually played backwards (although the sound runs forward to avoid overly disturbing interpretative possibilities). The second scene, S2, is a black and white scene depicting Leonard sitting in a motel room looking and feeling confused. His confusion is depicted through his voiceover narration. The third scene, S3, then returns to a color scene. It starts with Leonard pointing at Teddy’s picture to the receptionist at the motel counter, before Teddy shows up at the reception and walks to the motel garage with Leonard. The middle image of S3 in Figure 5 shows the long shot which depicts their walking from reception to garage. The long shot clearly shows the motel name *Discount Inn* on a big sign seen on the upper part of the screen. Leonard drives Teddy to an abandoned building where Teddy is then shot dead by Leonard. The narrative in this scene therefore directly precedes and overlaps with that of the first color sequence (S1). Finally, the second black-and-white scene (S4) continues Leonard voice-over narration from the previous black-and-white scene in the same motel room.

Drawing on the detailed cohesion analysis by (Tseng and Bateman, 2012) and (Tseng, 2013), the beginning four scenes of *Memento* are non-linear and have no clear spatio-temporal or logical relations across the color and the black and white scenes. Nevertheless, there are sufficient cohesive cues to interpret the characters and settings across the four scenes. Figure 6A summarizes the straightforward pattern of cohesive chains of the main characters and settings across the four scenes. As the chains show, Leonard and Teddy are both presented visually in S1. The *Leonard* chain shows that the reappearance of Leonard’s face is tracked in S2 and S3, while his name as “Lenny” was explicitly identified by Teddy in S3. The *Teddy* chain shows that Teddy is visually presented in S1, and his name is also explicitly written on his photo seen in the beginning shot in S3, before he appears in the motel counter. The chain of the first setting, the building, connects the visual repetition of the same setting in S1 and S3. This re-occurrence is further endorsed through the repetition of the same actions in the images where Leonard shoots Teddy. The chain of the second setting, the motel room, shows that the setting is

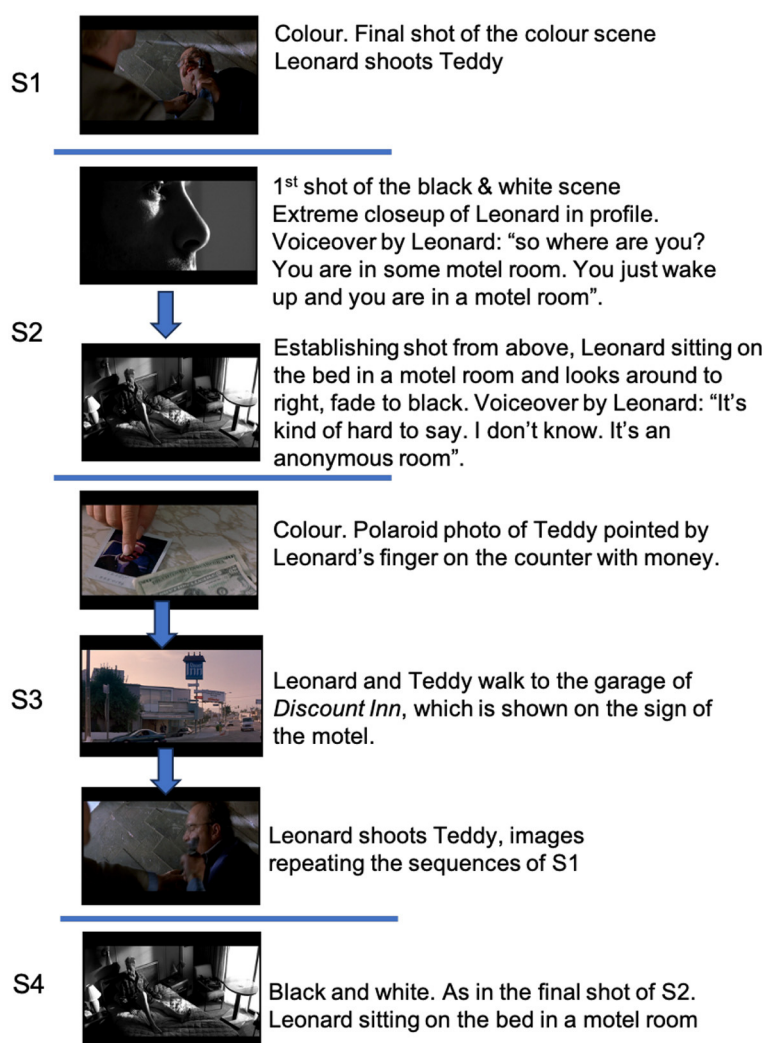


FIGURE 5  
Shots before and after scene transitions S1 to S4 in the opening sequence of *Memento*.

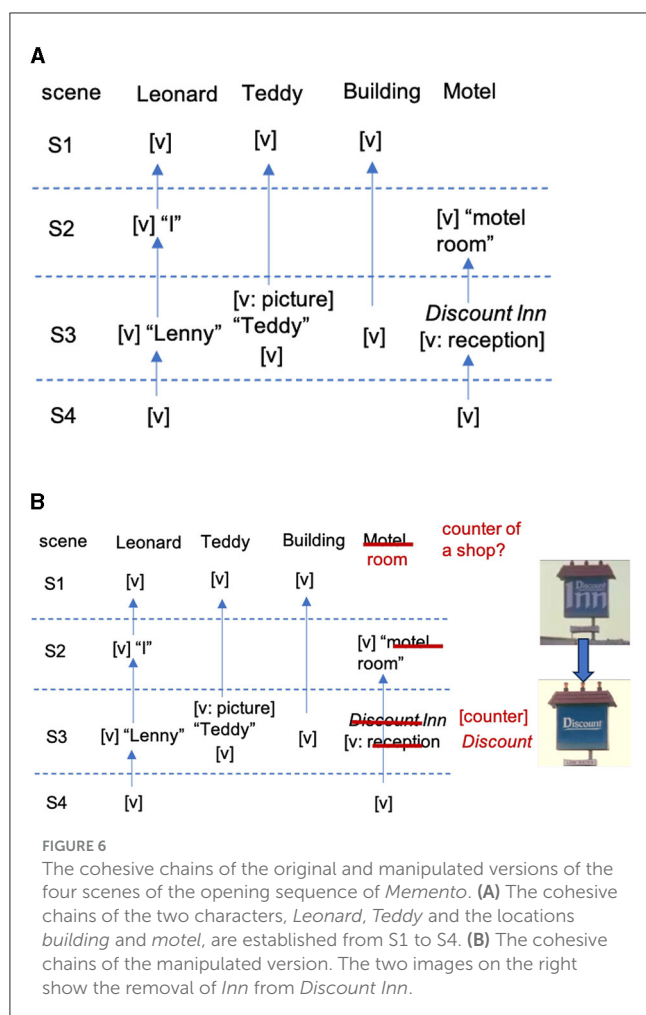
first explicitly identified as "motel room" in S2 by Leonard's spoken text "So you are in a motel room". In S3, the sign of *Discount Inn* and the visual image of motel reception then cohesively link back to the motel room in S2. There is no clear cue whether the room in S2 is in *Discount Inn* in S3 but the cohesive cue is nevertheless established through a *hyponymous* relation (i.e. motel room, motel reception and motel garage). The hyponymy could possibly lead the viewer to interpret the same setting. The same motel room explicitly reappears in S4.

Along the same lines of the previous study, we created the second version for comparison and manipulated the original sequence by removing the cohesive cues that direct the viewers to the specific interpretation of the second setting, the motel room. To this purpose, we wiped out the two lines of the spoken text by Leonard in S2, which explicitly refer the room as motel room: "You are in some motel room. You just wake up and you are in a motel room". That means, the viewers only hear Leonard saying "so where are you?" In S3, we also manipulated the motel sign, wiping out *Inn* from *Discount Inn*.

Figure 6B shows the cohesive analysis across the four scenes in the manipulated version. The main difference lies in the chain *motel room*. The original *motel room* chain now connects the generic room in S1 and S4. The scene setting in S3 is disconnected from the original motel chain to form an independent generic *Discount* shop counter chain, as any cue indicating the link between a shop/counter and a room is missing here.

Apart from manipulating cohesive cues for the two comparative sequences, namely, cued and uncued versions, we also manipulated the spatio-temporal order of the four scenes. As described above, the chronological sequence of the film story actually runs as S2-S4-S3-S1. In this chronological sequence, Leonard is in a motel room contemplating and plotting the murder of the assumed killer of his wife. He then emerges to the reception, encounters Teddy, and subsequently murders Teddy after driving from the motel to the building. Hence, we also prepared two versions of the scene orders, an original film version and the re-edited version with the S2-S4-S3-S1 order. Our hypothesis was that the re-edited sequence with chronological order might untangle the narrative complexity and





lead the viewers to interpret the correct event development, namely, Leonard was first in motel room in black and white scene, which precedes the color scenes. We also predicted that chronological order and cohesive cues impact each other in directing the viewers' scene comprehension. Cohesive cues may help the viewers to interpret scene order and vice versa, temporal cues of the scenes may improve the viewers' identification of the motel room setting, because the color scenes start with the location of reception counter and garage of a motel, which might increase the viewers' inferences of black and white scene as a motel room.

Hence, this study follows a 2x2 design, with cohesive cues and chronological scene order as two independent variables. The sequences, original and manipulated, were then presented to four different groups of participants and differences in their comprehension were measured.

### 3.2.1 Hypotheses

To assess the predicted effects, we test the specific hypotheses:

- Viewers of the manipulated, *uncued*, versions will be less certain about the specific identities of the motel room, even though the relevant visual elements inside the room are still readily accessible on screen.

- Viewers of the original, *achronological* version will be less certain about the event order of the black and white and the color scenes.

### 3.2.2 Experiments

The hypotheses were investigated by having participants evaluate one question and one statement immediately following their viewing of the beginning sequence of *Memento*:

1. Where is the setting of the black and white scene?
2. From the perspective of the characters, the black and white scenes happens before the color scenes.

As the previous study, the first one is an open question, while the second question was designed as a Likert scale. The viewer needs to select a response from the 5 points: 1 (totally disagree) to 5 (totally agree).

The comprehension test was conducted at the University of Bremen, and the participants ( $n = 74$ ) were undergraduate students who had not seen the film before the experiment. As the study had a 2x2 design, the participants were divided into four groups: Group 1 ( $n = 21$ ) watched the original achronological version without cohesive cues (with motel cues removed), group 2 ( $n = 17$ ) viewed the chronological version (edited s2-s4-s3-s1 sequence) with cohesive cues removed, group 3 ( $n = 17$ ) viewed the original achronological version with the cohesive motel cues, group 4 ( $n = 17$ ) watched the chronological version with cohesive cues.

### 3.2.3 Results

#### Question 1—Comprehension of the motel setting.

For analysing the open answers of the first question about the motel setting (*Where is the setting of the black and white scene?*), we coded the accurate answer (motel/hotel room) as 1 and all other answers as 0. Most inaccurate answers included "a room" or "sleeping room". In this study, we used dichotomous coding and treated any answer without mentioning "motel room" as incorrect. This indeed revealed a clear impact of cohesive cues. Nevertheless, as we will see in Study 3, we decided to refine the coding of the answer about the setting to finer gradations, which then uncovered more nuanced differences of participants' interpretation.

For the statistical analysis, we used logistic regression, suitable for modeling binary responses, to analyse the relationship between cohesive cues, temporal order and the viewers' comprehension. In general, the results show a significant effect of cohesive cues on the viewers' ability to establish the identity of the motel room ( $p = 0.0199$ ). The other independent variable, temporal cue, did not have a significant effect on the viewers' scene comprehension ( $p = 0.92763$ ).

More importantly, logistic regression analysis shows the relationship between cohesive cues and chronological order on the probability of correct comprehension of the motel room. This is demonstrated in terms of odds ratio—it was found that, holding chronological order constant, the odds of accurate comprehension decrease by 87% for the viewers who watch the sequence without cohesive cues, compared to the viewers who watch the sequence with cohesive cues. It was also found that, holding cohesive cues

constant, the odds of correct comprehension increase only by 5.3% for the viewers who watch the sequence with chronological temporal order, compared to the viewers who watch the sequence with the original, complex achronological order.

The above comparative result of odds ratio is visualized in the Figure 7. Here we can see the impacts of the two factors (“with” vs “without” cohesive cues, “chronological” vs “achronological” sequencing) to the correctness of participants’ answers.

A significant decrease of correct comprehension of motel room (between the probabilities of 1 and 0) if the cohesive cues are removed (namely, when data points move from “with” to “without” variable). In terms of “chronological and achronological” variable, there is no significant difference in the probabilities of correct comprehension. The two lines are nearly merged.

*Question 2—Temporal relation between black-white and color scenes*

For analysing the Likert scale results of question 2 (*From the perspective of the characters, the black and white scenes happens before the color scenes*), namely, about the temporal order of color and black and white scenes, we used the Align-and-Rank transform (ART) test. The results show a significance of cohesive cues ( $p = 0.0156$ ) in the viewers’ inferences of event orders across the scene transitions. The violin plot in Figure 8 shows the main difference of the two conditions (with and without cohesive cues). The distributions of the Likert scale score (1–5) for the two conditions are demonstrated through density curves - here we can see that in the original version with cohesive cues, a significant portion of participants is related to the score of 5 (totally agree), while the responses of the participants who watch the version without cohesive cues substantially vary, with more responses toward 1 (totally disagree).

However, the variable of temporal order did not have a significant effect on the viewers’ scene connections ( $p = 0.1379$ ). Moreover, no significant interaction effect between cohesive cues and temporal order was revealed ( $p = 0.7566$ ).

In summary, in this study, cohesive cues remain significant in leading the viewers’ interpretation of both the specific setting of the scenes (Hypothesis 1) and temporal order of event sequences (Hypothesis 2).

However, the second factor that we considered, the variable of chronological order of the scenes, does not have significant effects both on the comprehension of motel setting and on the interpretation of event sequence orders.

The reason for the weak effect of the second factor, the chronological order of scenes, might be attributed to the fact that *Memento* is a puzzle film characterized by Nolan’s signature complex, non-linear film structure. Although in the manipulated version, we tried re-ordering the four scenes to match its general story order (S2-S4-S3-S1), the scene transition between S4 and S3, namely, the black and white scene of Leonard in a motel room and the next color scene of Leonard at the motel reception, still exhibit a substantial ellipsis. This deliberate narrative gap in Nolan’s famous puzzle film might be the reason why the effect of the variable chronological order is diluted.

To refine our test design in order to further investigate the significance of chronological orders of scene relations theorized by (Bateman, 2007) and (Bateman and Schmidt, 2012), we conducted a

third study, using a sequence of a more straightforward drama film, *Julie & Julia*. With this film material, in the next study, we were able to refine our experimental measures and broaden our questions to include the viewers’ confidence level of their comprehension and their interpretation of the main characters’ intention. The rationale behind testing participants’ level of confidence in their own inferences was to test whether the manipulation has resulted in any uncertainties among participants as to their own judgements regarding the setting and the goal of the main character. Testing participants’ self-rated level of confidence could provide insight into whether there was a discernible difference in the perceived confidence influenced by the manipulation.

### 3.3 Study 3: Julie & Julia

As described above, in the third study, we tested the same independent variables, cohesive cues and temporal orders, but we refined our measures and used a sequence extracted from a non-puzzle film. The sequence also deals with three transitions across four scenes.

Figure 9 presents four representative shots from each scene. Shot 1 depicts the first scene in the living room of the main character, Julie Powell, while she is seen typing on her laptop and reading aloud to her husband a passage from her blog, wherein she recounts her unsuccessful cooking attempt from the previous day. Shot 2 shows the second scene, in which she walks on the street before entering a butcher’s shop. In shot 3, Julie has entered the shop. Inside the shop, Julie is seen purchasing ingredients for one of the recipes she is attempting from Child’s cookbook. In this scene, the shop setting is filled with conventional visual cues of a butcher’s shop, e.g. meat displayed behind the counter. The viewers can also hear Julie off-screen voiceover depicting her cooking plan throughout shot 2 and 3. This scene is then cut to the kitchen setting, shown in shot 4, where Julie is already back from the butcher’s shop and is cooking using the ingredients she just purchased from the shop.

While creating film materials for all experimental conditions, we decided to first remove the overall spoken text of the entire sequence (including the dialogues between Julie and her husband and Julie’s voiceover). The reason is that the spoken text is highly indicative for Julie’s plan to go to the butcher’s shop and purchase meat in the shop. We wiped out the entire spoken text to remove verbal cohesive cues leading to the setting of the butcher’s shop. Nevertheless, removing the entire verbal text for the version without cohesive cues could lead to the substantially loose control of two conditions because we also wiped out other verbal information that is relevant to the overall narrative interpretation. Hence, to secure clean effects through the experimental control, we decided to remove and replace the spoken text with the film’s soundtrack music for all conditions first and then manipulate visual cohesive cues and temporal orders based on these sequences already without verbal text.

Figure 10A illustrates the analysis of cohesive chains of the version with visual cohesive cues. As we can see in the chain pattern here, no spoken verbal cues are included, as they were all removed.

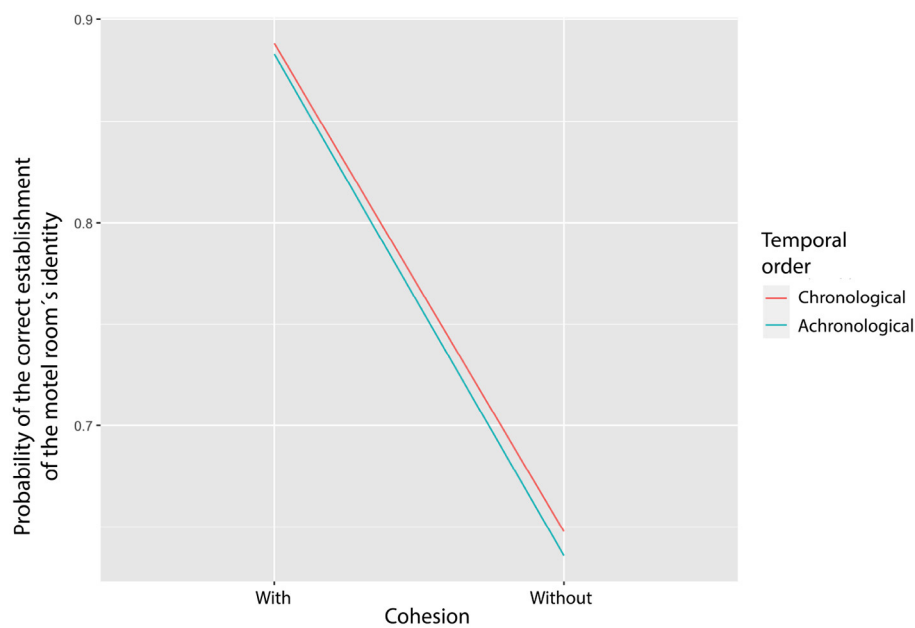


FIGURE 7

The probability of the correct establishment of the motel room identity across the groups with/without cohesion and chronological/achronological order of the scenes.

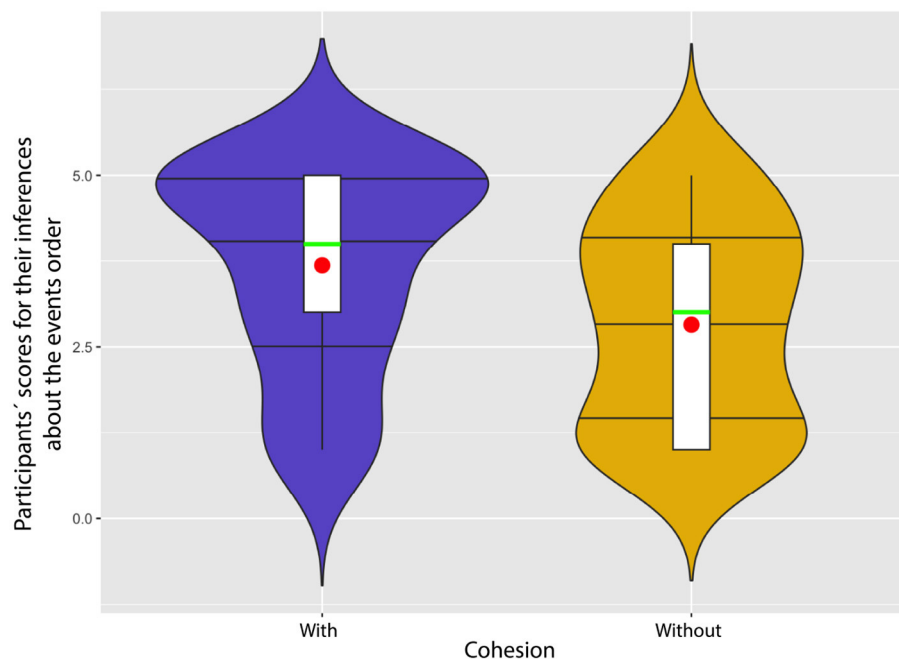


FIGURE 8

Main effect of cohesive cues (with and without) on participants' inferences of the events ordering.

Across the four scenes, three reoccurring narrative elements are tracked. The first *Julie* chain shows her reoccurring appearance throughout the entire segment. The second chain is the setting of Julie's *home*, it is first presented with a living room setting and is cohesively linked through a hyponymous relation by another home setting, the kitchen. The third chain is a butcher's shop. It is

presented in shot 2, the specific identity of the *butcher's shop* chain is introduced through textual visual cues such as *K&T Quality Meats, Meat and Poultry* and the associated price tags on the window. In shot 3, visual cues such as butcher's outfit, meat and cheese products inside the shop cohesively link the setting to the *butcher's shop* chain introduced in the previous shot.



FIGURE 9  
Selected shots in *Julie & Julia*.

The events across the scene transitions can then be depicted based on the chain pattern as follows:

*Julie is first at home and then she goes to the butcher's shop, before returning home again.*

To manipulate visual cohesive cues, we targeted the setting of the butcher's shop. We removed all visual cohesive cues from the butcher's setting (shop front door in shot 2 and indoor setting in shot 3) that reveals the specific identity of the shop. Figure 11 shows exactly what visual cues were removed from the original scenes. Here we can see that the text *K&T Quality Meats, Meat and Poultry* and the price tags for the meat products are written on the roof and the windows of the shop that Julie is entering in shot 2. In the uncued version, these texts have been removed and we can see that Julie is entering an indoor space with no written indication of its identity (Figure 11A). Moreover, in the version with cohesive visual cues we can see Julie talking to the butcher who is attired in a traditional white butcher's costume (Figure 11B), placing and weighing meat pieces on the scale. There are also refrigerators stocked with jars, meat and cheese, big chunks of cheese hanging off the ceiling and price tags on the counter's glass. Contrasting this, in the uncued version (bottom image of Figure 11B) we turned the butcher's conventional outfit into a blue shirt and a red hat. We have also removed all food products, price tags, the scale and the meat pieces in the butcher's hand, so that it is no longer identifiable what he is putting on the counter.

Figure 10B illustrates the cohesive analysis across the four scenes in the version without cohesive cues. Similar to the previous two *Memento* studies, the removal of cohesive cues transforms the specific shop identity (here the butcher's shop) into a generic, non-specified indoor space. Hence, the same setting chain now includes only visual

elements of a generic indoor space, such as a door and a counter.

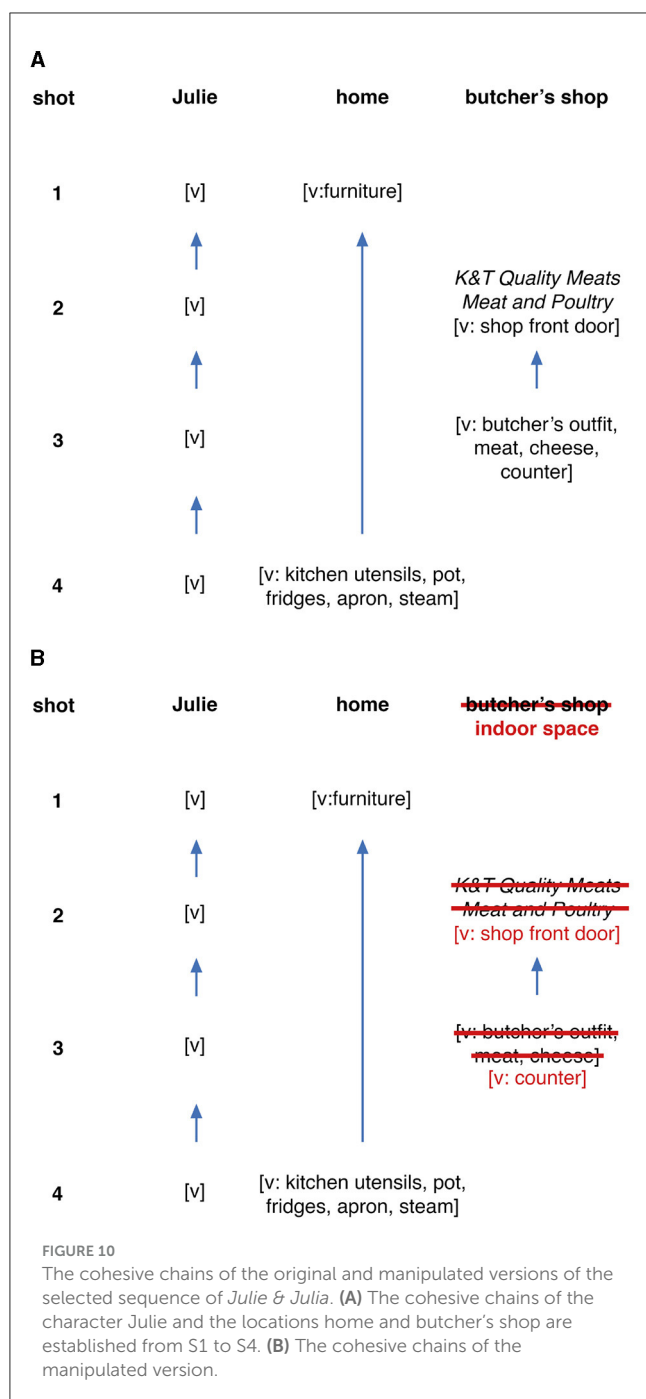
In addition to the removal of the cohesive visual cues and for testing the second factor, namely, the temporal scene order, we re-edited the temporal relation between the four scenes. That means, based on the original temporal order (S1-S2-S3-S4), we edited the order into an alternative one, S4-S1-S2-S3. In this alternative version, viewers first see Julie cooking in the kitchen, followed by the living room scene and subsequently, Julie's visit to the butcher's shop. This was done in order to temporally change the logical relation between the scene in the butcher's shop and the scenes at home. Our hypothesis following the scene order modification is that, the home cooking scene (S4) directly followed by Julie's visit to a shop (S3) might lead the viewers to infer the setting as a food-related shop, even in the absence of the cohesive visual cues of butcher's shop.

### 3.4 Hypotheses

To assess the predicted effects of cohesive cues and temporal orders, we tested the specific hypotheses:

- Viewers of the uncued (without cohesive visual cues) and alternative temporal order version will be less accurate at establishing the identity of the butcher's shop.
- Viewers of the uncued (without cohesive visual cues) and alternative temporal order will be less accurate at identifying the goal and intention of Julie's actions in the story.
- Viewers of the uncued (without cohesive visual cues) and alternative temporal order version will be less confident in their inferences about the butcher's shop identity and the goal of the female character's actions.





- There will be a difference in participants' time perception of the story events between the original and a orders of the segment.

### 3.5 Experiments

The experiment was conducted at the University of Bremen. All participants ( $n = 76$ ) were students or employees of the university who had not seen the movie before. Each participant was allocated to one of the four experimental groups: Group 1

( $n = 19$ ) watched the original temporal order version with cohesive cues, Group 2 ( $n = 19$ ) watched the original temporal order version without cohesive cues, Group 3 ( $n = 19$ ) watched the alternative temporal order version with cohesive cues, Group 4 ( $n = 19$ ) watched alternative temporal order version without cohesive cues.

For a general comprehension check, that is, to make sure that participants were also able to follow the non-manipulated part of the excerpt, we also asked them some basic comprehension questions to reveal whether they noticed that there were 3 characters in total, that Julie was typing and talking in S1 and that the husband was holding a bike. Hence, we first asked the following questions:

- How many characters are in the video clip (both main and secondary)?
- What was the female character doing when she was talking to the man in the living room?
- What was the man holding when he was leaving the living room?

The questions of comprehension check were immediately followed by the questions listed below to address the hypotheses of cohesive cues and temporal order:

1. In what kind of place was the female character when she was talking to the other man?
2. Why do you think she went there? Be specific.
3. How sure are you about your answer to the previous question?
4. Estimate approximate time period shown in the video clip.

To avoid leading language, we used the wording "the other man" when referring to the butcher. This decision was based on the arrangement of questions in the questionnaire provided to participants. The question directly followed two questions that mentioned "the man in the living room".

For analysing the answers, participants' responses to question 1 and question 2 were converted into points that ranged from 0 to 3, where 0 indicated the least accurate answer, and 3 represented the right establishment of the butcher's shop or the goal of the female character going there. Participants who explicitly stated that the female character went to the designated place to buy ingredients received a score of 3. Examples of such responses include "She was buying ingredients for cooking" and "She probably wanted to buy some ingredients for dinner". Those who inferred that she was buying food were awarded a score of 2, as can be seen in the following response "She went there to get lunch". Participants who mentioned shopping, without specifying ingredients or food, received a score of 1, for example, as in the response "womöglich um sich irgendwas zu kaufen" (perhaps to buy something). Those who did not mention any of the above received a score of 0, as indicated in the response "to pick up some parcel".

Question 3 is about participants' confidence in their previous responses. It is estimated using a five-point Likert-Scale question, ranging from 1 (not sure) to 5 (very sure).

Question 4 addresses participants' time perception of the story events. Participants were given the following options: two hours, four hours, one day, more than one day.



FIGURE 11

Screenshots from the original and manipulated versions of the butcher's shop scene in *Julie & Julia*. (A) Scene outside of the butcher shop: The version with cohesive visual cues (top) versus the version without cohesive visual cues (bottom) in *Julie & Julia*. (B) Scene inside of the butcher shop: The version with cohesive visual cues (top) versus the version without cohesive visual cues (bottom) in *Julie & Julia*.

### 3.6 Results

#### General comprehension check

The general comprehension check shows that participants across the four groups understood the overall, non-manipulated part of the sequence.

#### Question 1 - Comprehension of the setting identity of the butcher's shop

For analysing responses of this question, we used the ART test. In general, the main effects of both independent variables, cohesive cues and temporal order, were significant.

In terms of the variable of cohesive cue, the analysis results show differences in participants' comprehension of the segment between the version with and without cohesive cues. In the conditions where cohesive cues were present, participants were significantly more accurate at establishing the identity of the butcher's shop. This result is visualized in Figure 12A. Here one can see that participants from the condition with cohesive cues on average received higher scores than participants in the condition without cohesive cues. [Mean ( $M$ ) = 2.553, Interquartile Range ( $IQR$ ) = 1] compared to the uncued conditions, where visual cues were absent ( $M$  = 1.684,  $IQR$  = 0), as revealed by the Align-and-Rank transform (ART) test ( $p < 0.05$ ). The plot shape indicates greater variation in participants' responses in the condition without cohesive cues, while those in the conditions with cohesive cues exhibited higher agreement.

Unlike the previous study on the puzzle film *Memento*, we found that the temporal order in this study played a significant role in comprehending the butcher's shop setting. This result is

shown in Figure 12B in which we can tell the difference that participants who watched the manipulated, alternative temporal order version on average received higher scores ( $M$  = 2.237,  $IQR$  = 1) than those who watched original temporal order version ( $M$  = 2,  $IQR$  = 0), as shown by the ART test ( $p$  = 0.038) (Figure 12B). As described above, we predict that the alternative temporal order version which brings the home cooking scene before the butcher shop scene enhances the inferences of the shop as a food/cooking relevant shop. The plot shape also reveals that in the original temporal order conditions, the majority of participants received the score of 2. In contrast, the alternative temporal order conditions exhibited a more wide spread distribution between scores 2 and 3, with more participants receiving a score of 3.

With regard to the interaction effect between the visual cues and the temporal order, our results show no significance, as the interaction effect between the two factors tested in Study 2.

#### Question 2 - Main character's intention in the story

To analyse participants' responses of question 2, namely, "Why do you think she went there?", the ART test revealed that the removal of cohesive cues indeed led to a deterioration in participants' ability to comprehend the goal of Julie in the segment ( $p$  = 0.013). As shown in Figure 13, the average score is significantly higher in the conditions with cohesive cues ( $M$  = 2.132,  $IQR$  = 0.75) compared to the conditions without cohesive cues ( $M$  = 1.632,  $IQR$  = 1). The  $IQR$  illustrated in the plot indicates that, in the conditions without cohesive cues, the middle 50% of participants received scores below 2, while in cued conditions, the middle 50% received scores above 2.

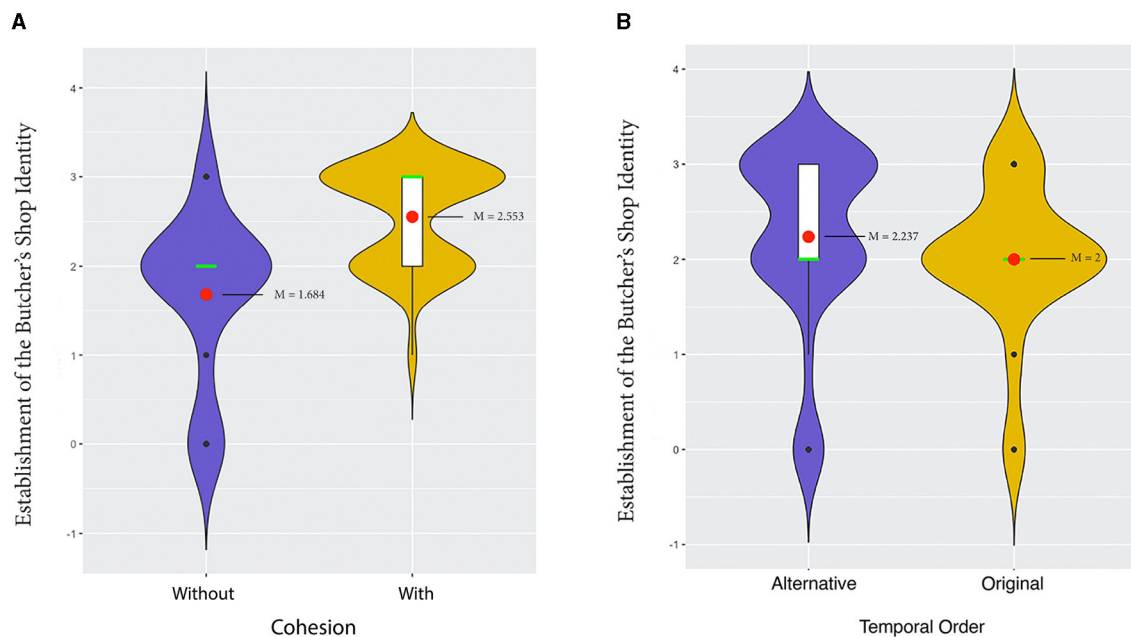


FIGURE 12

Participants' scores for establishing the identity of the butcher's shop based on the presence of cohesive cues (without/with) and temporal order (alternative/original) in *Julie & Julia*. (A) Main effect of cohesive cues on participants' ability to establish the butcher's shop identity. (B) Main effect of temporal order on participants' ability to establish the butcher's shop identity.

In terms of the second factor, the temporal order, the effect was not significant ( $p > 0.05$ ). Our analysis also reveals that there was no significant interaction effect between cohesive visual cues and temporal order.

#### Question 3 - Confidence level of viewers in their own responses

At first glance at the results, we note that none of the participants who watched the extract without cohesive cues rated their confidence at level 5 (representing "very sure"). Similarly, none of the participants who viewed the extract with visual cues rated their confidence at level 1 (indicating "very unsure") in both the establishment of the butcher's shop identity and the establishment of the causal relation.

We then used two-way ANOVA to analyse responses to question 3. Our results show that the main effect of cohesive visual cues on confidence level was highly significant ( $p < 0.001$ ). As illustrated in Figure 14, the comparative findings indicate that participants in conditions with cohesive cues displayed significantly higher confidence in their inferences ( $M = 3.5$ ,  $SD = 1.033$ ) compared to participants in conditions without cohesive cues ( $M = 2.474$ ,  $SD = 0.893$ ).

As in the analysis of the questions 1 and 2, the effect of temporal order on confidence level was not statistically significant ( $p > 0.05$ ) and there was no significant interaction effect between visual cohesive cues and temporal order either.

#### Question 4 - Viewers' time perception of story events

For analysing the responses of question 4, we used the ART test again to test the effect of cohesive cues and temporal scene order on the time perception of the story events.

A significant main effect of temporal order was observed ( $p = 0.024$ ), indicating that participants estimated the approximate

duration of the events taking place in the story to be *longer* in the conditions with original temporal order ( $M = 2.105$ ,  $IQR = 0.75$ ) compared to conditions with alternative temporal order ( $M = 1.737$ ,  $IQR = 1$ ), as illustrated in Figure 15. The IQR illustrated in the plot indicates that, in the alternative order conditions, the middle 50% of participants rated that the events in the sequence took less than four hours. In contrast, in the original temporal order conditions, the middle 50% of participants estimated the events took more than four hours.

There was no significant interaction effect between the independent variables on participants' temporal perception of the segment ( $p > 0.05$ ). The results show that the main effect of cohesive cues on time perception was not statistically significant ( $p > 0.05$ ).

## 4 Discussion and conclusion

The results of our three studies have further empirically supported our hypothesis that cohesion in film is highly relevant and significant in people's comprehension of scenes and settings whether during a continuous scene or transition across different scenes and whether in a complex puzzle film or in a narratively straightforward film. We also provide results showing that cohesion is significant in viewers' inferences of character's intention in the story. Moreover, we also show the significance of temporal order of scenes in viewers' inferences of both scenes and settings and the length of event time.

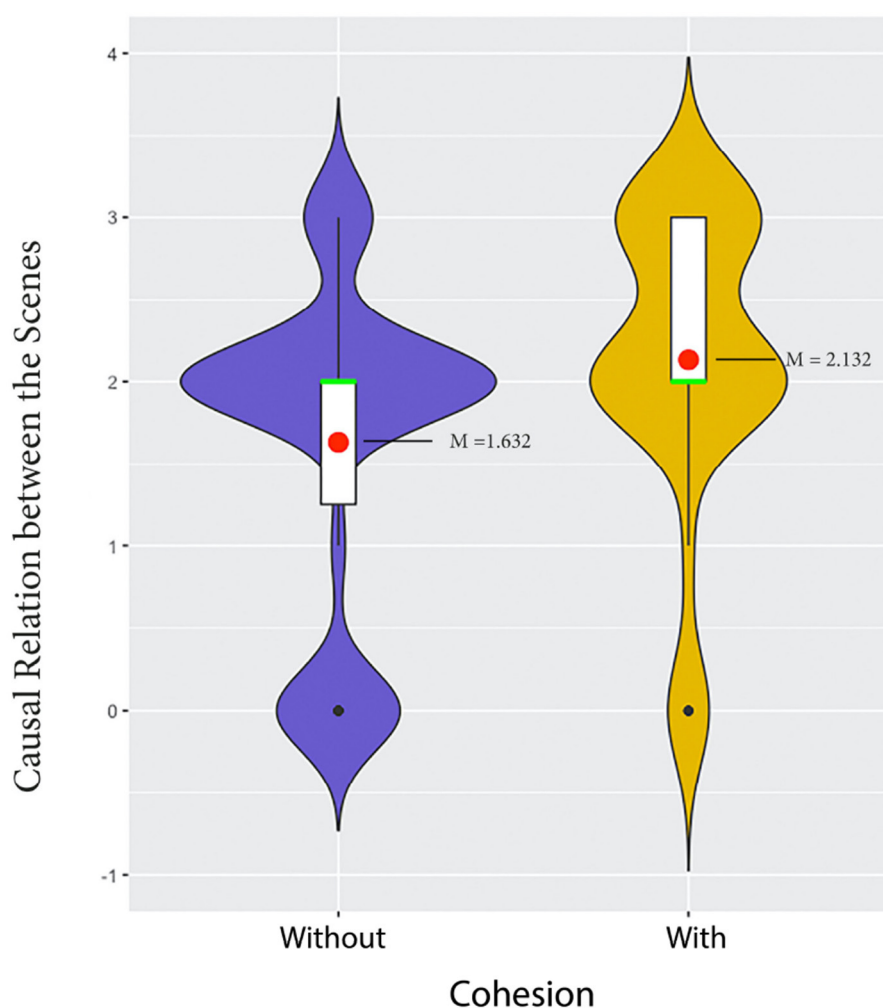


FIGURE 13

Main effect of cohesive cues on participants' ability to establish the goal of the main character in *Julie & Julia*.

The empirical results of our study lead to more questions and hypotheses for future investigation.

First of all, the cultural background of viewers might impact their understanding of a film. For instance, two participants in our Study 3 noted after the experiment that such butcher's shops were more prevalent in the countries of their origin than in Germany. This observation is similar to the previous research on the role of cultural background in narrative comprehension (Horiba, 1990), which demonstrated that, when reading about scenes taking place in Japan, native Japanese readers focus on more intricate details and utilize their cultural knowledge about the local details to infer the protagonists' actions. In contrast, non-native Japanese readers did not exhibit the same degree of event details in their narrative comprehension. Hence, we believe that cultural origin of viewers could be a relevant factor for different ways of establishing cohesion within a scene and could be a crucial variable to investigate empirically.

Another question to dive deeper into is what narrative features impact viewers' interpretation of intentions and goals of characters.

The event comprehension model (Zacks, 2007) proposes that the changes of space, time and intention all lead to the comprehension of event change. However, there has not been sufficient research indicating whether these factors actually interact—our study 3 shows that space (cohesive cues of setting) is significant in viewers' comprehension of character's intention, while time is not a significant factor for story intention. Hence, more empirical tests are thus required to untangle the inter-relation of these factors for event comprehension.

Our study 3 also explores the intriguing issue of time perception and its relation to space in film. Our results indicate that the difference in the perception of the temporal length of events in the two experimental groups (original and alternative versions) actually rests on the different narrative spatial structures. For instance, in the non-chronological version, in which S4 in Julie's kitchen is edited directly before S1 in Julie's living room, the story event time seem shortened for the viewers due to the contiguous space relation between S4 (kitchen) and S1 (living room). We hence further hypothesize that event perception of two contiguous spaces



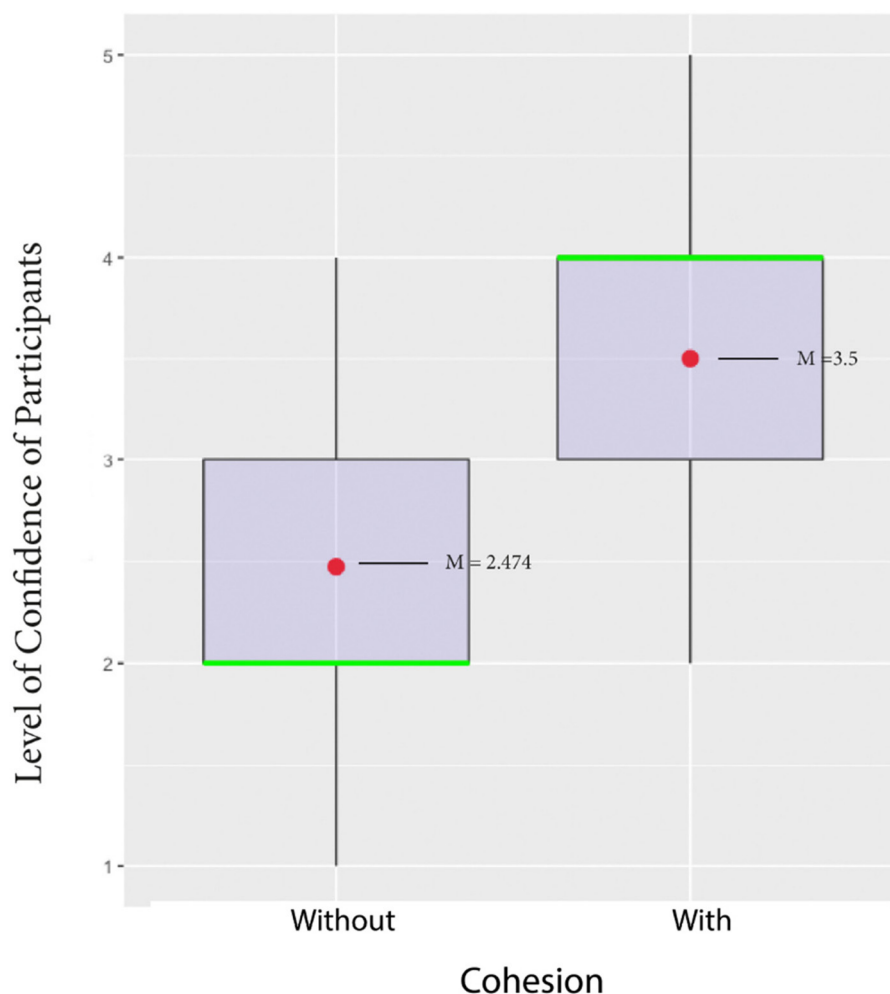


FIGURE 14

Main effect of cohesive cues on participants' level of confidence in their inferences in *Julie & Julia*.

could also lead to the event interpretation of closeness in event time. The hypothesis will require further empirical studies.

In this paper, we have presented results of three empirical studies conducted with the aim of investigating how multimodal cohesion in film influences viewers' narrative comprehension of events across scene transitions. While the previous research (Tseng et al., 2021) has indicated that the absence of cohesive cues leads to an uncertainty about the setting within a continuous scene, we have broadened the test scope about multimodal cohesion in three aspects: (1) we tested how cohesive cues function to carry viewers across scene transitions (study 1), (2) we added another factor, namely, temporal order of scenes theorized by Bateman (2007), to investigate how these two factors impact narrative comprehension independently and interactively, (3) apart from testing viewers' correct understanding of setting identities, we also tested the viewers' confidential level about their answers, whether their understanding of character's intention and event time perception are related to the two factors, cohesive cues and temporal scene orders.

We also identify limitations of conducting experiments using cinematic materials. It is challenging to predict if a film material offers enough control of stimuli. The refinement of our experimental from Study 1 to Study 3 shows our endeavor to shift from *Memento* to *Julie & Julia* in order to extend the questions that can be addressed in a more controlled fashion.

We hope our empirical studies on multimodal cohesion demonstrate a valuable combination of empirical methods and multimodal discourse analysis, which is a robust, textual-based model highly valuable for investigating people's cognitive processes through uncovering how people maximize coherence when perceiving multimodal artifacts. Finally, we also hope to have shown how the multimodal film research endeavors in the last decade by Bateman (2007), Bateman and Schmidt (2012), Tseng and Bateman (2012), and Tseng et al. (2021) continue to develop and shed light on significant aspects of human perception and meaning interpretation of film narratives.

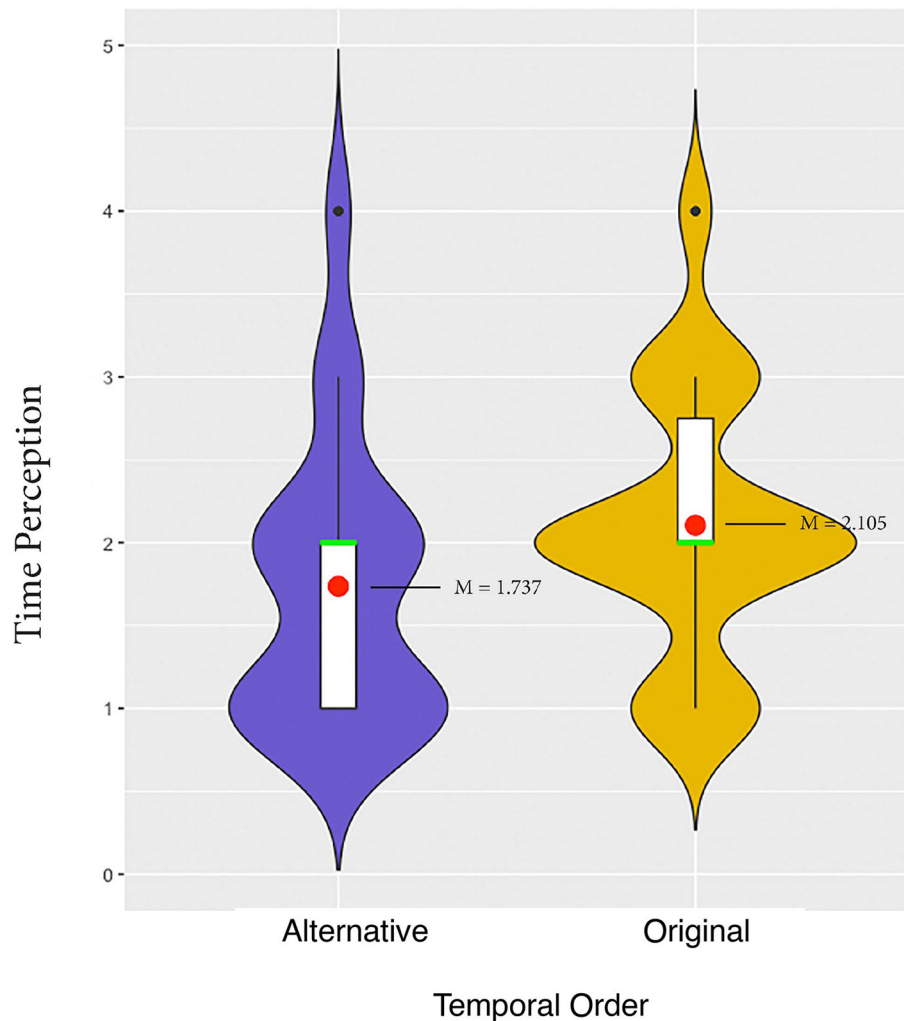


FIGURE 15  
Main effect of temporal order on participants' time perception in *Julie & Julia*.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

Ethical approval was not required for the studies involving humans because in our studies, participants watched movie clips and then completed a questionnaire to measure their understanding. No personal data was collected. Participants were informed about any potential risks and their participation was strictly voluntary, with the option to withdraw at any time. All collected data was kept strictly confidential. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of data included in this article.

## Author contributions

DM: Writing – original draft. C-IT: Writing – original draft.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The research in this paper was conducted with the support of the University of Bremen.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

- Bateman, J. A. (2007). Towards a grande paradigmatique of film: Christian Metz reloaded. *Semiotica* 167, 13–64.
- Bateman, J. A., and Schmidt, K.-H. (2012). *Multimodal Film Analysis: How Films Mean*. Routledge Studies in Multimodality. London: Routledge.
- Bateman, J. A., and Tseng, C. (2013). The establishment of interpretative expectations in film. *Rev. Cognit. Linguist.* 11, 353–368. doi: 10.1075/rcl.11.2.09bat
- Bordwell, D. (2008). *The Hook: Scene Transitions in Classical Cinema*. Online Essay (David Bordwell's Website on Cinema) Available online at: <http://www.davidbordwell.net/essays/hook.php> (accessed March 15, 2024).
- Drummond, T., and Wildfeuer, J. (2020). "The multimodal annotation of gender differences in contemporary tv series. Annotations in scholarly editions and research," in *Functions, Differentiations, Systematization*, 35–58.
- Eisenstein, S. (1969). *Film Form: Essays in Film Theory*. Eugene: Harvest Book.
- Ephron, N. (2009). "Julie & Julia," in *Columbia Pictures, USA (film)*.
- Halliday, M. A. K., and Hasan, R. (1976). *Cohesion in English*. London: Longman.
- Halliday, M. A. K., and Matthiessen, C. M. I. M. (2013). *Halliday's Introduction to Functional Grammar*. London and New York: Routledge.
- Hasan, R. (1984). "Coherence and cohesive harmony," in *Understanding Reading Comprehension: Cognition, Language, and the Structure of Prose*, ed. J. Flood (Newark, Delaware: International Reading Association), 181–219.
- Horiba, Y. (1990). Narrative comprehension processes: a study of native and non-native readers of Japanese. *Modern Lang J.* 74, 188–202.
- Loschky, L. C., Larson, A. M., Magliano, J. P., and Smith, T. J. (2015a). What would jaws do? The tyranny of film and the relationship between gaze and higher-level narrative film comprehension. *PLoS ONE*. 10, e142474. doi: 10.1371/journal.pone.0142474
- Loschky, L. C., Ringer, R. V., Ellis, K., and Hansen, B. C. (2015b). Comparing rapid scene categorization of aerial and terrestrial views: a new perspective on scene gist. *J. Vision* 15, 1–29. doi: 10.1167/15.6.11
- Martin, J. R. (1992). *English Text: Systems and Structure*. Amsterdam: Benjamins.
- Metz, C. (1974). *Film Language: a Semiotics of the Cinema*. Oxford and Chicago: Oxford University Press and Chicago University Press.
- Nolan, C. E. (2000). "Memento," in *Summit Entertainment, USA (film)*.
- Smith, T. J. (2012). The attentional theory of cinematic continuity. *Projections*. 6, 1–27. doi: 10.3167/proj.2012.060102
- Tseng, C. (2008). "Cohesive harmony in filmic text," in *Multimodal Semiotics: Functional Analysis in Contexts of Education*, ed. L. Unsworth (London: Continuum), 87–104.
- Tseng, C. (2012). Audiovisual texture in scene transition. *Semiotica* 192, 123–160.
- Tseng, C. (2013). *Cohesion in Film: Tracking Film Elements*. Basingstoke: Palgrave Macmillan.
- Tseng, C., and Bateman, J. A. (2012). Multimodal narrative construction in christopher Nolan's Memento: a description of method. *J. Visual Commun.* 11, 91–119. doi: 10.1177/1470357211424691
- Tseng, C., and Bateman, J. A. (2018). Cohesion in comics and graphic novels: an empirical comparative approach to transmedia adaptation in city of glass. *Adaptation* 11, 122–143.
- Tseng, C., Laubrock, J., and Pflaeging, J. (2018). "Character developments in comics and graphic novels: a systematic analytical scheme," in *Empirical Comics Research: Digital, Multimodal, and Cognitive Methods*, eds. J. L. Alexandra Dunst, J. Wildfeuer, and A. Dunst (London: Routledge).
- Tseng, C.-I., Laubrock, J., and Bateman, J. A. (2021). The impact of multimodal cohesion on attention and interpretation in film. *Discourse, Context Media* 44, 100544. doi: 10.1016/j.dcm.2021.100544
- Tseng, C.-I., and Thiele, L. (2022). Actions and digital empathy in interactive storytelling of serious games: Multimodal discourse approach. *Soc. Semiot.* doi: 10.1080/10350330.2022.2128039
- Zacks, J. (2015). *Flicker: Your Brain on Movie*. Oxford: Oxford University Press.
- Zacks, J. M., Speer, N., Swallow, K., Braver, T., and Reynolds, J. (2007). Event perception: a mind/brain perspective. *Psychol. Bull.* 133, 273–293. doi: 10.1037/0033-2909.133.2.273



## OPEN ACCESS

## EDITED BY

Janina Wildfeuer,  
University of Groningen, Netherlands

## REVIEWED BY

Elisabeth Zima,  
University of Freiburg, Germany  
Jana Holsanova,  
Lund University, Sweden

## \*CORRESPONDENCE

Leandra Thiele  
✉ le\_th@uni-bremen.de

RECEIVED 15 December 2023

ACCEPTED 18 March 2024

PUBLISHED 11 April 2024

## CITATION

Thiele L, Schmidt-Borcherding F and  
Bateman JA (2024) All eyes on the signal? -  
Mapping cohesive discourse structures with  
eye-tracking data of explanation videos.  
*Front. Commun.* 9:1356495.  
doi: 10.3389/fcomm.2024.1356495

## COPYRIGHT

© 2024 Thiele, Schmidt-Borcherding and  
Bateman. This is an open-access article  
distributed under the terms of the [Creative  
Commons Attribution License \(CC BY\)](#). The  
use, distribution or reproduction in other  
forums is permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original publication in  
this journal is cited, in accordance with  
accepted academic practice. No use,  
distribution or reproduction is permitted  
which does not comply with these terms.

# All eyes on the signal? - Mapping cohesive discourse structures with eye-tracking data of explanation videos

Leandra Thiele<sup>1\*</sup>, Florian Schmidt-Borcherding<sup>2</sup> and  
John A. Bateman<sup>1</sup>

<sup>1</sup>Linguistics and Literary Studies, Faculty 10, University of Bremen, Bremen, Germany, <sup>2</sup>Pedagogy and Educational Sciences, Faculty 12, University of Bremen, Bremen, Germany

In this paper, we consider the issue of how the fine-grained multimodal design of educational explanation videos, such as those widely available on YouTube and other platforms, may be made accessible to empirical studies of reception and effectiveness. This is necessary because previous research has often led to conflicting conclusions concerning the roles of particular design elements. We argue that this may largely be due to insufficient characterizations of multimodal design itself. To achieve tighter control of this potential source of variation, we present a multimodal descriptive annotation framework drawing on multimodal (cohesive) film discourse analysis. This framework is seen as a critical first step toward being able to highlight just those differences in design that have functional consequences. For such consequences to accrue, however, viewers need to attend differently to corresponding design differences. The goal of the current paper, therefore, is to use eye-tracking techniques to explore the extent to which discourse structures revealed by our analytic framework relate to recipients' attention allocation. We hypothesize that any potentially emerging anomalies in regards to discourse organization, such as instances of unsuccessful cohesion signaling, may have correlations in the behavioral data. We report our current state of development for performing this kind of multimodal cohesion analysis and some of the unresolved challenges raised when considering how such analyses may be related to performance data.

## KEYWORDS

cohesion, multimodal cohesion, discourse analysis, eye-tracking, education videos, explanation videos, multimodality

## 1 Introduction

Explanation videos are now extremely popular in both informal and formal educational settings. They draw on different disciplines and areas of knowledge and appear in many different forms, such as short videos, "Reels," and so on, each potentially exhibiting substantial differences in design. Explanation videos are also available on-demand on many online platforms (e.g., YouTube), which played an important role in furthering their use and acceptance during the COVID-19 pandemic worldwide (cf., e.g., [Yaacob and Saad, 2020](#); [Breslyn and Green, 2022](#); [Trabelsi et al., 2022](#); [Lu, 2023](#)). However, even before this latest explosion in use, there was already a long established research tradition addressing the question of what makes an explanation video effective (or not). A host of principles and guidelines have been proposed (e.g., [Mayer, 2021b](#)), but empirical results often paint a far more mixed picture see the discussion below and further references



in, for example: (Bateman and Schmidt-Borcherding, 2018; Bateman et al., 2021). We see here substantial methodological issues that need clarification before attempting to gauge effectiveness. Unless we are able to characterize differences in design in a manner that reveals precisely not only which design differences may have functional consequences but also the conditions under which such consequences are most likely to follow, it is unlikely that consistent empirical results will be obtained. In short: it is important to be able to distinguish between mere physical differences in design that may have little effect on viewers' engagement and the differences that play an active role in interpretation-building, for better or worse.

In this paper we propose a methodology that enables us to focus specifically on this challenge of isolating differences in design that have measurable behavioral consequences. We consider this as an essential step prior to being able to conduct more reliable and discriminating effectiveness studies. In order to ascertain whether particular differences in design correlate with reception differences, we employ results of an eye-tracking study to examine the extent to which correlations can be found between the gaze data and our proposal for a fine-grained account of the discourse structure of explanation videos. This may then help to develop further hypotheses concerning discourse structures and those structures' contribution to the achievement of communicative goals, particularly the goals of effectively informing and explaining. Our focus here, however, will be solely on our discourse analysis framework and the support it offers for mapping cohesive structures with eye-tracking data, leaving the final question of the role that such structures may play for effectiveness for subsequent studies.

We see this intermediate step as essential because of what is probably the primary challenge raised by attempting to deal systematically with data of this kind: that is, the highly multimodal nature of explanation videos. Such videos avail themselves of the full range of expressive forms now supported by the medium and so readily combine diverse kinds of broadly "written" representations, such as written language, graphs, tables, and mathematical formulae, more pictorial, schematic, or diagrammatic representations, as well as "second-order" visual resources for navigation and other purposes, such as circles, underlining, arrows, boxes, overall layout and other segmentation techniques. This has made any characterization of "design" in a manner sufficiently precise to be supportive of empirical investigation a major bottleneck for further inquiry.

In this paper we focus on an approach capable of addressing this issue in a general manner by drawing on contemporary linguistically-inspired theories of multimodality. The term "multimodality" refers to the phenomenon of combining multiple semiotic modes, i.e. different ways of representing meaning, in the service of coherent communication. This would seem ideally suited to the complex multiple expressive forms found in explanation videos. However, traditional conceptions of semiotic modes relying on broad labels such as "written text," "image," "sound," etc. have often conflated formal and functional properties making them difficult to apply in research. For example, the

functions served by "words" in diagrams, graphs, pictures, or captions can be, and often are, quite different, which makes ascertaining their contribution to design (or lack of it) challenging. Similarly, the functions played by visual materials, such as diagrams, mathematical equations, or graphs can only be found in combination with the other deployed resources they co-occur with.

To resolve these difficulties, we adopt the position and methods for multimodality research specifically argued in Bateman et al. (2017). This account offers a more formalized account of multimodal communication that assumes a tight connection between expressive forms and the discourse functions of those forms regardless of presentation modality just as is required to handle the multimodal complexity of explanation videos. In addition, the descriptive framework we set out is intended to be strongly supportive of corpus-based work on explanation videos, by means of which we can more effectively triangulate between descriptions, empirical reception studies, and diverse medial realizations.

The paper builds on an earlier exploratory study by Bateman et al. (2021), in which the feasibility and utility of a rich multimodal annotation scheme for capturing the interplay of different semiotic modes in explanation videos was demonstrated. We now develop this scheme further and report on a pilot empirical investigation seeking correlations between the multimodal annotation scheme and recipient data collected for contrasting videos using eye-tracking methods. By these means we support our claim that the broader task of relating fine-grained design choices to video effectiveness may usefully be broken down into several components: here we focus specifically on a first stage of relating design to performance data which may subsequently, as suggested above, be brought more readily into contact with learning effects.

We structure the paper as follows. We begin in Section 2 with a brief review of the state of the art among those approaches that attempt to gain more analytic hold on factors responsible for educational videos being effective or not. The diversity of results found in these studies supports our basic claim that we need to tease apart the factors contributing to design more finely. To assist the development of such studies in the future, we report in Section 3 on the multimodal annotation scheme that we are developing for explanation videos and the specific functional aspects of discourse that are currently covered. Although our annotation scheme is defined to apply to explanation videos in general, for the purposes of the present paper we need also to be sufficiently detailed as to show micro-scale interactions with behavioral data. Consequently, in Section 4, we draw on results of an eye-tracking study carried out for some explanation videos that were specifically constructed to exhibit controlled variation, focusing on the gaze data gathered with respect to one particularly complex slide. This allows us in Section 5 to begin to address our central question—that is, the extent to which theoretically motivated multimodal annotations, and in particular the discourse structures revealed, can be shown to correlate (or not) with behavioral data. This results in several challenges and directions for extending the multimodal annotation in future work that we summarize in Section 6. Finally, in Section 7, we summarize what has been achieved and the goals established for the next steps to be taken in such studies.

## 2 Brief literature review and state of the art

Research on the instructional effectiveness of educational videos dates back far beyond contemporary platforms such as YouTube, beginning in the 1970s (cf. Bétrancourt and Benetos, 2018). Since then considerable attention has been paid to potential relations between the effectiveness of videos and their design. Within the domain of instructional design the most prominent and most recent theoretical research contexts for educational videos are given by cognitive and perceptual multimedia learning frameworks, e.g., Cognitive Load Theory (CLT: Paas and Sweller, 2021), Cognitive Theory of Multimedia Learning (CTML: Mayer, 2021a), and Integrative Text and Picture Comprehension (ITPC: Schnotz, 2021).

Many recommendations for instructional design or principles of multimedia learning that are derived from these theoretical frameworks (for a comprehensive overview, see Fiorella and Mayer, 2021) can be considered to hold for educational videos as well (Fiorella, 2021). Some of these principles are almost naturally fulfilled due to technical characteristics of videos. For example, the multimedia (Mayer, 2021b) and multiple representations (Ainsworth, 2021) principles pronounce that using verbal and visual modes fosters learning compared to relying on a single (re)presentation mode. The modality principle (Castro-Alonso and Sweller, 2021) states that once verbal and visual modes are used in combination, verbal information should be spoken rather than written. Obviously, most educational videos consist of (non-verbal) visualization accompanied by (spoken) text. Other design principles, however, need to be actively addressed when creating an educational video. For example, the simultaneous use of multiple visual representations forces a split of visual attention that should either be avoided as much as possible (Ayres and Sweller, 2021), or be supported by additional signals that guide a learner's (visual) attention (van Gog, 2021). Still other principles may lead to ambiguous or even contradictory interpretations of an actual video design. For example, in educational videos the instructor can be visible in the video as a "talking head" or only audible as a "voice over" (e.g., Wang and Antonenko, 2017). However, on the one hand, visible instructors are a source of split attention, whereas, on the other hand, they may serve as a social cue (Fiorella and Mayer, 2021). Indeed, a recent review of the effects of instructor presence in instructional videos found mixed results (Henderson and Schroeder, 2021).

Theoretical explanations for most of these principles, as offered by the above mentioned theories of multimedia learning, often assume the mental integration of multiple different external representation types, mainly verbal and pictorial representations, the perception of these representations via different sensory modalities, mainly the eye and the ear, and different verbal and visuo-spatial mental representations of information (e.g., Mayer, 2014; Schnotz, 2014). While psychological and psycholinguistic research has achieved some consensus on models of verbal or "propositional" representation (Kintsch and van Dijk, 1978; Kintsch, 1988), models of visual information are separated across different kinds of visualizations such as graphs (e.g., Kosslyn, 1993), pictures (e.g., Levin et al., 1987), diagrams (e.g., Larkin and Simon,

1987; Tversky et al., 2000), or animations (e.g., Tversky et al., 2002; Ainsworth, 2008). This may be one reason why models of text-picture-integration stay incomplete or underspecified in their understanding of the mental integration process itself. As Bucher and Niemann (2012, 292) note, it is important to clearly separate distinct visual representations in both form and function. Largely "pre-theoretical" distinctions such as that commonly made between "words" and "image" do not provide sufficient discrimination since both written language and images are visual and both are commonly integrated in a range of distinct semiotic contexts. Moreover, the presentation of information in such materials continuously makes references between (spoken) language and visualization as well (e.g., in "verbal signaling").

Also relevant here is largely independent work carried out on academic presentations because these often overlap significantly with the kinds of presentations found in many explanation videos. Whereas, strictly speaking, explanation videos form a broader class, whenever those videos employ presentations of the form found in academic presentations using software such as PowerPoint, there are useful empirical results obtained in that domain to build upon (Schnettler and Knoblauch, 2007; Wiebe et al., 2007; Bucher et al., 2010; Bucher and Niemann, 2012), as well as proposals for the multimodal description of such presentations (e.g., Rowley-Jolivet, 2004). All such approaches point to the need to provide finer-grained accounts so that variations in reception and effect may be investigated more closely.

There are, moreover, interesting differences and similarities to consider between work on live presentations, using tools such as PowerPoint, and the medial variants found in explanation videos. Whereas researchers increasingly study the role of gestural signaling of relevant information during a presentation (Bucher and Niemann, 2012), explanation videos commonly employ visual signaling that is designed into the material of the visual presentation by means of graphical highlighting with arrows, areas of color, and so on, often animated. These can be expected to play a particularly important role whenever presenters are not visually present. Here, there remains much to consider, relating, for example, diagrams and gestures more closely, as proposed by Tversky et al. (2013) and Kang et al. (2015), as well as empirical and descriptive work on infographics (Habel and Acartürk, 2006; Martin and Unsworth, 2023).

For the present paper particularly relevant are then findings in cognitive studies that propose the signaling or cueing principle, which suggests a higher learning outcome from multimedia learning resources when those incorporate certain signals to guide viewers "to the relevant elements of the material or [to] highlight the organization of the essential material" (van Gog 2021; see also Richter et al., 2016; Schneider et al., 2018; Alpizar et al., 2020, and Mayer 2021b). Ozcelik et al. (2009) put forth two concrete hypotheses related to this, namely the *guiding attention hypothesis*, which suggests relevant information receive more attention when given signaling as well, and the *unnecessary visual search hypothesis*, which refers to the ease of locating related information between visual and verbal modes. What these then have in common is a lack of crucial information concerning precisely *what* these signals are that guide viewer attention and how they have to be meaningfully woven into any material's organization. In this regard,

these principles need further refinement on an empirical basis as now attempted in several lines of research (Richter et al., 2016; Mayer et al., 2020).

Finally, it is interesting to note that there has been surprisingly little work to date attempting to relate aspects of cohesion, the specific functional discourse phenomenon we employ below, and eye-tracking data, even with purely verbal texts. From the multimodal perspective relevant here, for example, Acartürk et al. (2014) report on a study of the effects of different styles of cross-references to figures in a constructed “text”-“figure” composite layout. Although systematic differences in gaze behavior (particularly durations for attending to the text and to the visual figure) were found, the layouts of the stimuli used were highly unnatural and did not reflect the multimodal complexity of the kinds of data considered here. The lack of natural stimuli for such experiments is a common difficulty that our provision of fine-grained annotations for design is also intended to alleviate.

In the subsequent sections we propose a contribution to the goals of theoretical and practical refinement by utilizing the far more fine-grained characterization of the possibilities for signaling and guiding viewer attention offered by multimodal analysis. This will allow us to investigate to what extent signaling as realized at the design level in video data is consistent with empirically measured viewing behavior. By these means we aim for an additional empirically-supported “filter” capable of focusing analytic attention on just those features of design that may be critical for subsequent uptake; whether or not that uptake has consequences for the *effectiveness* of an explanation must then be subject to investigation in its own right.

### 3 An annotation framework for explanation videos

In this section, we introduce our general multimodal-descriptive annotation framework for explanation videos. The purpose of this framework is to support fine-grained investigation of the discourse structures of educational videos, which we hypothesize play a central role in guiding those videos’ reception. More specifically, we show how a multimodal discourse analysis may capture aspects of “textual” organizations corresponding to the signaling principles introduced above; by these means signaling principles in multimodal discourse receive a concrete realization that we can then subject to empirical analysis.

We first introduce the overall organization of the annotation scheme and its practical realization within the annotation software ELAN, developed at the Max-Planck Institute in Nijmegen (Wittenburg et al., 2006; ELAN, 2023). This scheme draws on and extends the account first motivated and introduced in Bateman et al. (2021). We then explain its use for one specific area of multimodal discourse organization, that of multimodal cohesion. This is the area that we will use below when exploring potential correlations with eye-tracking behavior. We will draw examples of the annotation scheme in use from the eye-tracking analysis that we perform below, although the scheme itself is intended quite generally for characterizing communication of this kind.

### 3.1 Annotation of complex audiovisual data

In order to move toward multimodal analysis that is sufficiently fine-grained to support empirical study, Bateman and Schmidt-Borcherding (2018) argue that the precise discourse placement of mobilized expressive resources in any kind of medium participating in discourse may be critical. The discourse structures involved therefore need to be captured so that organizational “weak spots” may be identified. In the current case, we will seek to operationalize such potential weak spots in terms of multimodal cohesion. As remarked in the introduction, multimodal discourse structures may be expressed using a rich diversity of representational forms, including various kinds of written and iconic representations as well as “second-order” visual resources for navigation and showing text organization. These all need to find a place in the developed analysis and annotation scheme.

Following annotation techniques long established in linguistic corpus work and since extended for multimodal corpora as well (cf. Bateman, 2013; Knight and Adolphs, 2020), the rich diversity of information required is captured in the annotation scheme by means of defining multiple layers of distinct kinds of information. Many studies of multimodal phenomena adopt broadly similar “layered” schemes of data annotation of this kind. In our case, however, we draw additionally on the more specific guidelines for multimodal corpus work set out in Bateman (2022), whereby distinct kinds of information are made to correspond broadly to individual semiotic modes, including all aspects of the formal definition of semiotic modes argued by Bateman et al. (2017). It is the systematic application of this definition that begins to establish a “meta-language” for comparing and contrasting explanation videos in general.

Currently, the modes included in the annotation framework and considered in our analyses are the following:

- verbal speech
- written language
- diagrams
- graphs
- mathematical formulae
- tables and corresponding tabular elements (i.e., columns and cells and labels)
- arrows and lines (sometimes within diagrams, sometimes not) that function representationally with respect to the information being presented
- highlights (including, e.g., arrows/lines/circles, speech bubbles, etc.) that function textually to orchestrate engagement with the information being presented.

Substantial interaction can be found between all of these forms of expression. However, for present purposes, we will focus discussion primarily on aspects that have been found particularly challenging up to now, such as the interplay of the *visual* elements. This is by no means to be taken as suggesting any prioritization of the relevance of distinct modes over others.

Methodologically, the fact that we are working with a temporally-based medium allows the levels of description to be linked back to the original data by timestamps. This makes it

appropriate to model these layers as “tiers” in ELAN. Thus, each tier of information segments an analyzed video temporally with respect to some specified facets of the video’s multimodal organization. Explanation videos often deploy further media as part of their information presentation that may then provide additional spatiotemporal structuring of their own which must also be captured. One common medium used in this way is Microsoft’s PowerPoint or similar tools; these media are “slide”-centered, which we then treat similarly to scenes in more film-like videos. In all cases, it is the perceptible visual material that is considered for analysis not the production—that is, if a slide develops by introducing animated elements that might be implemented in separate slides but which appear continuous, then these are treated as a single temporally unfolding unit. Establishing properly motivated analytic units of this kind is an essential step for reliable analysis (cf. Bateman et al., 2017, 2021).

Most of the individual forms of expression to be included are captured as “base” layers of annotations. These offer a foundation for defining several further kinds of multimodal annotation that are essential for capturing inter-relations and signaling techniques operating between elements expressed in different semiotic modes. *Relational* information of this kind is often not supported by current annotation tools, ELAN included. To handle this in a general fashion, we have developed annotation guidelines for including relational properties that build on existing annotation tool capabilities. These guidelines then also stand as a method supporting the use of ELAN for multimodal data whenever richly internally structured multimodal ensembles are involved.

For present purposes, relational information is mainly needed for two types of tiers. First, arrows, lines, circles, and speech bubbles generally relate to other elements in the videos—arrows and circles for example, commonly serve to highlight other elements, whereas lines connect elements. This information needs to be captured in addition to the bare presentation durations given by the segments of the base-level tiers and, moreover, can well require their own duration information—for example, an arrow intended to draw attention to some other element may appear and disappear independently of the durations of the elements being referred to.

Analytic units with their own durations are most commonly represented in ELAN and similar time-based annotation tools as layers or tiers in their own right. Consequently, in our framework, relational units also all receive their own tiers within the ELAN annotation. Information about the temporal extent of an element’s visibility (or audio duration for verbal speech) is then given by defining time interval segments within these tiers as usual, marking the respective starting and ending times of their occurrence. For ease of reference, these tiers are labeled following a specific naming scheme identifying the *structural* position of any component within the presentation as a whole. Thus, for example, an ELAN tier label `data-point:6_d:1_s:4` picks out the sixth visual “data point” that is part of the first diagram (`d:1`) of slide 4 (`s:4`).

The relational information itself, i.e., the relations between these elements and the units they relate, is then captured using structured labels stored directly as annotation values of the relevant interval segments of the base tiers. These structured labels identify both those further elements that the marked elements relate and the type of linkage, currently either highlight, connect, or label. Figure 1 shows as an example three tiers whose elements

either highlight or connect with other elements. Those other elements are identified throughout by their respective tier names as just described. Thus: the first line of the figure captures the information that a particular circle within the first diagram on slide 4 (`circle:5_d:1_s:4`) functions to highlight a particular data point in that diagram (`data-point:6_d:1_s:4`), which will also have its own independent tier elsewhere in the annotation. Connection relations are given similarly by mentioning both elements being related.

Information also needs to be given concerning the form of these relational elements, e.g., the colors of circles, lines, etc. Although it would be possible formally to add such information to the structured labels just introduced, this would lead to potentially very complex interval annotations that ELAN provides no support for and which would likely become increasingly error-prone. Thus, rather than over complicate the information maintained in the interval labels, we instead employ ELAN’s “Comments” functionality for recording visual properties directly. Examples are shown in Figure 2. Here we see that annotations in the comment section are also linked to specific time stamps allowing properties to be anchored to time intervals as well—this would be needed when, for example, the color or shape of an arrow or some other unit changes during its use. This type of information is annotated for all semiotic modes whenever relevant. Thus a further example would be when the color or forms of textual elements change; this then also includes, as we shall see below with respect to form-based cohesive signaling, form properties for numbers in math formulae or written language.

The annotation scheme described so far then provides most of what is needed for engaging with the rich multimodality of temporally-based complex media such as explanation videos. Data sets annotated in this way would offer a strong foundation for investigation of the use that is being made of the resources captured by the application of several methods, such as, for example, corpus-based studies and, as pursued here, behavioral measurements and experimentation.

## 3.2 Multimodal cohesion in explanation videos

For addressing the particular use of multimodal resources for signaling and guiding interpreters, we now turn to the notion of multimodal cohesion, as this is generally taken as one of the primary techniques by which texts, of any kind, provide additional interpretation cues for their recipients. Cohesion as adopted and refined here was originally defined by Halliday and Hasan (1976) solely with respect to verbal language. Cohesion is said to be active whenever elements of a text require interpretations of other elements of the same text in order to receive their own interpretation: most prototypical examples of this would be pronouns, where the interpretation depends on identifying their intended referents. Relatively early in work on multimodality this notion of cohesion was extended to apply to “texts” consisting of more than verbal language. Royce (1998), for example, set out a system of several distinct kinds of “cohesive” relations operating across written texts and accompanying images and diagrams. The



circle:5_d:1_s:4 [1]	highlights: data-point:6_d:1_s:4
arrow:1_d:1_s:4 [1]	connects: data-point:6_d:1_s:4,mean-value-age:0_d:1_s:4
arrow:2_d:1_s:4 [1]	connects: data-point:6_d:1_s:4,mean-value-words:0_d:1_s:4

FIGURE 1  
ELAN in-tier inter-relation annotations showing the use of structured annotation “labels” rather than terms selected from controlled vocabularies or free text.

Start Time	End Time	Tier	Comment
00:04:42.783	00:05:49.162	circle:5_d:1_s:4	color:black
00:04:45.398	00:05:49.162	arrow:1_d:1_s:4	color:green
00:04:48.086	00:06:13.692	arrow:2_d:1_s:4	color:green
00:05:01.283	00:05:48.907	Math-formula:1_pt...	color:green

FIGURE 2  
Additional visual annotation information recorded in ELAN comment sections.

function of such connections was to suggest explanations for how texts could guide recipients to bring together different sources of information, each potentially expressed with different semiotic modes. Several extensions of this basic idea have been proposed since; [Liu and O’Halloran \(2009\)](#) provide a detailed overview as well as some significant further proposals of their own that we will also draw on below.

Many accounts offered of multimodal cohesion to date have followed Royce’s lead in focusing on “text-image” relations. As we have seen above, however, this would not be appropriate for explanation videos as a far broader diversity of semiotic modes are usually at work. The underlying theory for the analytic steps we implement here are consequently based more on the audiovisually extended framework of multimodal cohesion analysis developed by [Tseng \(2013\)](#). This method calls for the construction of cohesive chains for audiovisual data regardless of the semiotic modes employed. Elements are linked in cohesive chains when they stand in particular discourse relations, such as co-referentiality as mentioned for pronouns above. Cohesion analyses are then shown using cohesive chain diagrams which depict the re-occurrence relations active in a text. This allows, in the multimodal case, the combined use of semiotic modes to be shown in a structured way so that the various contributions of multimodal resources can be tracked exhaustively across a text’s development.

Several quite specific extensions to the notion of cohesion inherited from its application for verbal texts need to be made for the multimodal context, even for the treatment of verbal language. One of these concerns the fact that in any multimodal artifact, there may be several units realizing verbal language co-present, both spatially and temporally. This means that some of the basic distinctions for cohesive analysis need to be refined. Co-referential cohesion in traditional verbal language, for example, is typically distinguished according to the “direction” of the relating cohesive link. More specifically, the relationships of situational identity constructed by co-referentiality across a text can occur in two ways: either the relationship is prospective, termed *cataphora*, or retrospective, termed *anaphora*. Anaphors thus “look back” to

their referents, while cataphors “look forward.” In single linearly organized “monomodal” texts these two directions naturally exhaust the possibilities as two referring expressions may always be ordered with respect to one another. However this does not hold for multimodal communication since multiple “contributions” may co-exist, co-occurring at the same time (matching on temporality) across different or multiple instances of the same modes. We add this third kind of referential cohesion to our account and term it “*co-phoric*.” We propose that multimodal referential cohesion may contribute to recognition of many of the signaling principles mentioned above and so may play a role in guiding a viewer’s attention, which should in turn leave behavioral traces, such as differences in gaze behavior as we investigate below.

A further source of potential cohesive ties when considered multimodally relates to the forms of the deployed expressive elements rather than their referents. When, for example, various elements co-present in a video are related by selecting particular colors, then this may serve as a signaling device calling for recipients to bring together the identified elements in some way, but not requiring that those elements be seen as co-referential. This form of connection is relatively under-researched in the context of accounts of cohesion, although clearly of importance for design. Both “intersemiotic parallel structures” and “intersemiotic parallelism” in [Liu and O’Halloran \(2009\)](#) account might be extended to include this.

We now include these forms of cohesion explicitly in our annotation scheme as they may clearly play an important role for discourse coherence. The way in which form information, such as shape and color, is captured in the annotation was already described above (cf. [Figure 2](#)). This technique is then also used to cover referential cohesion as follows. First, co-referentiality information is annotated directly in a distinct type of ELAN tier labeled as “cohesive links.” Intervals defined within these tiers then “pick up” elements from specified base tiers that are related to other modes through co-referentiality. Thus, for verbal speech, for example, co-referential items in other semiotic modes are linked to the

respective verbal elements by entries in a corresponding “verbal-speech-cohesive-links” tier. The cohesive links tier thus identifies annotations of linguistic verbal tokens that are co-referenced in the discourse across modes. The corresponding information concerning the tiers to be linked to these tokens, i.e., the referents, is again given in the ELAN’s comment section as shown in Figure 3. These annotations have the specific structure “coref: [linked tier name].” Thus, the first line of the figure captures the fact that there is a verbal element (whose contents is captured in the corresponding interval segment defined in the verbal-speech tier) which is co-referential with another element in the video, the data point labeled `data-point:6_d:1_s:4`.

Since in the multimodal case co-reference can occur across any semiotic modes capable of referring, and not just the verbal, we generalize this method to allow co-referential information for any tiers describing any semiotic modes by similarly adding corresponding cohesive link tiers. These then operate in the same way as for the verbal cohesive links tier, simply picking out non-verbal elements as required. Thus, although the verbal speech track in our medium of investigation generally provides a good orientation for engaging with all of the other material presented, this is not necessarily the case. Nevertheless, for our present object of analysis, it is often appropriate to select the verbal mode as the main axis of discursive organization and development as we shall see.

Finally, as Liu and O’Halloran (2009) emphasize, cohesion analyses of all kinds can be seen from two perspectives: a static, product-oriented perspective (the “synoptic” view) and a dynamic, text-development perspective (the “logogenetic” view). The cohesion analyses that we will mostly present in this paper are synoptic in the sense that they do not reflect the temporal development of the audiovisual “texts.” This raises significant questions when engaging with the *reception* of these texts since this clearly occurs over time. How these may be related in empirical work will then be an important topic we take up below.

### 3.3 Multimodal cohesion diagrams

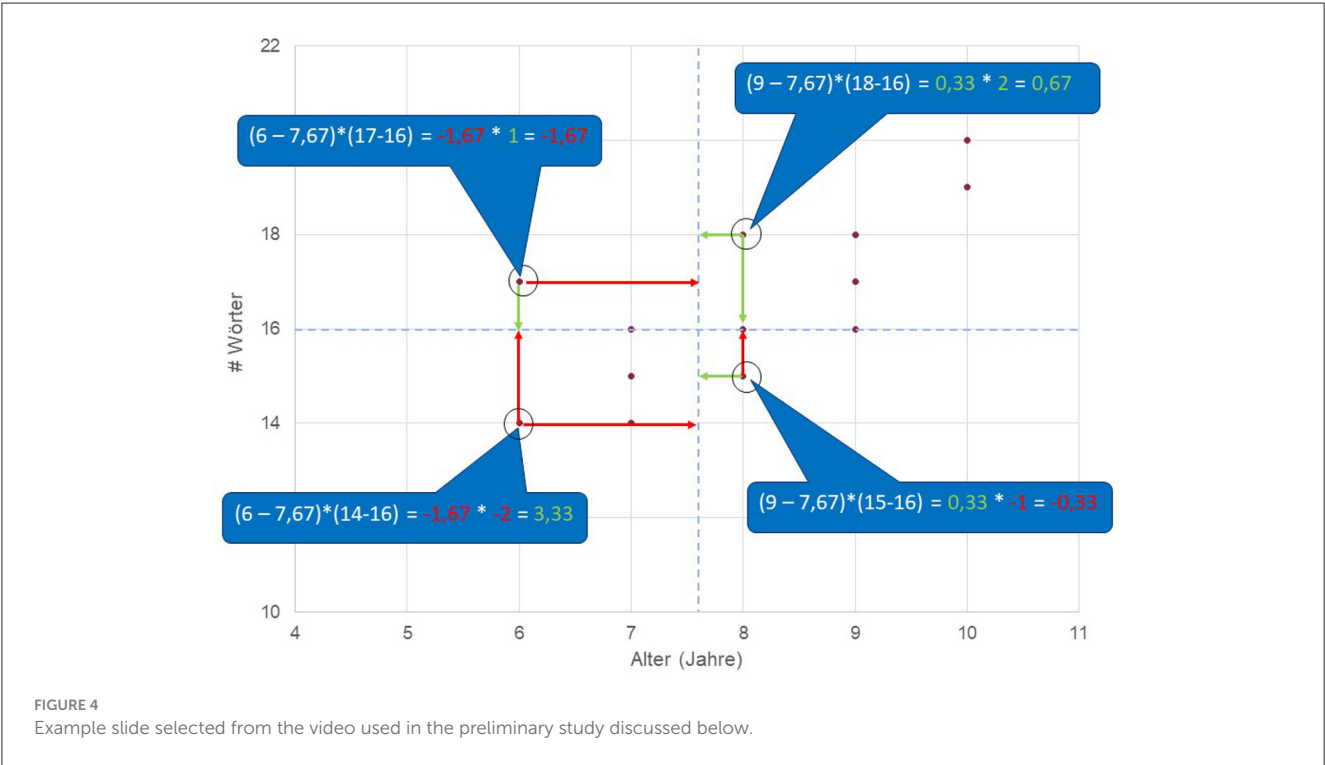
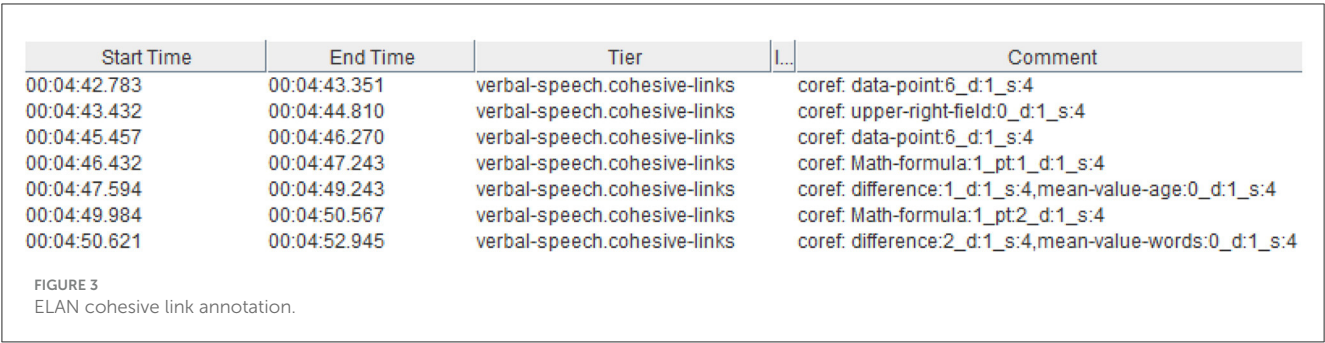
When constructing and inspecting cohesion analyses, it is traditional to use visualization diagrams where identified cohesive chains—i.e., elements in the analyzed texts that are connected cohesively—are shown running vertically down the page with cohesive links between the elements of single chains depicted by vertical arrows. Thus, for example, repeated mentions of a particular data point, first with a full referring expression such as ‘the data point’ and subsequently by various forms of pronominalization, would all be placed in a single cohesive chain running vertically down the page. Whereas in Tseng (2013) these cohesive chains might already combine expressions in various semiotic modes, for example including a graphical data point as well, for current purposes we begin by separating the cohesive chains across semiotic modes. This is intended to allow us to focus more specifically on the work that recipients have to do in finding relationships across the various forms of expression used and is also motivated by the sheer diversity of modes that we need to separate analytically. Thus, in our case, a verbal reference

to a particular graphically depicted data point would involve two cohesive chains: one for the verbal language and one for the visual. These two chains are then linked by, in this case, a co-referentiality relation. In our visualizations, such relationships are depicted by arrows running horizontally across the page connecting the tiers concerned.

Most earlier visualizations of multimodal cohesion analyses have been constructed manually, which quickly becomes difficult when the multimodal complexity of an analyzed text increases. Now, since we have a formally specified annotation scheme for capturing multimodal cohesion, we generate such chain diagrams automatically using a specifically written R script running directly on the ELAN exported data. We will make extensive use of cohesion diagrams below when comparing audiovisual discourse structures with our eye-tracking data, and so it will be useful here to show a worked example in detail. For this, we take a single slide used in the explanation videos that served as stimuli in the eye-tracking experiment we draw upon; this is also the slide that we focus on below. In addition, this visual presentation is accompanied by a verbal track describing how the graphic is to be interpreted and this verbal information is naturally also a necessary component of the cohesion analysis.

Figure 4 shows the slide in question. The videos themselves were made in German for a German-speaking audience, and so all the written text labels visible in the slides and the accompanying spoken language are in German as well. The overall topic of the presentation used in the video is “Covariance and Correlation,” discussing how values measuring these statistics are calculated with respect to data. The screenshot in the figure correspondingly depicts the final state of a slide where this calculation is explained by the lecturer working through a concrete example in which the ages of a set of 15 children are placed in relation to their respective active vocabularies. The data used for the calculation of co-variance being discussed in the example is depicted visually by means of data points positioned on a graph. The “age” of the respective children is shown running along the horizontal “x”-axis and their respective vocabulary sizes (“words”) run vertically on the “y”-axis. The task of the narrator/presenter in the video at this point is to lead the students through some selected data points (each point depicting a particular child) so as to make clear the respective relationships between the information concerning individual children and the average age and vocabulary size for the group as a whole. The instructor’s verbal description is shown transcribed with standard Jefferson notation (e.g., Jefferson, 2004) along with a simple English gloss on the left of Figure 5. The precise calculation to be learned by the students is given in the mathematical formulae picked out by the four call-outs in the slide resembling “speech balloons” from comics. This is itself an interesting case of the influence of the design of the “semiotic software” used, in this case PowerPoint, as such speech balloons are included in the graphic resources readily on offer (cf. Zhao et al., 2014; Djonov and van Leeuwen, 2022), but may well then be employed for purposes other than denoting speech.

The slide is naturally quite complex in its own right, consisting of the data points, the overall graph within which the data points are placed, visual representations of differences between x and y values and group averages, visual highlights of individual points, and call-outs showing the mathematical formulae required to perform the



calculations necessary for four selected data points. There are many questions concerning how to present such information effectively, both visually and in combination with the verbal description. The static depiction of the entire slide as it appears in Figure 4 corresponds to only one (and clearly not the best) of many possible presentational styles that would be possible in the dynamic medium of an actual explanation video. For example, the presentation might be aided by a more gradual build-up of the information on display. This is precisely the dimension of variation that we return to specifically in the studies reported below.

For the purposes of establishing a synoptic, complete cohesion analysis, however, we simply need to characterize all of the units present in the visual field, the spoken language, the relationships among these, and the temporal extents over which these contributions unfold. Thus, even a synoptic representation automatically includes time because the materiality of the medium (specifically its *canvas*: Bateman et al., 2017) is inherently dynamic, making temporal extents a necessary component of its description. This means that time is included in the analysis,

but as an unchanging and unchangeable fourth dimension (i.e., a “block universe” view of time). This can then serve as a stable basis for subsequent analyses where the dynamic nature of textual unfolding may be explicitly considered more from the perspective of recipients rather than from the ‘product’ as a whole.

The multimodal annotation of this slide is then also correspondingly complex but remains nevertheless fully conformant with the framework introduced above. Indeed, the fact that we can now deal with this degree of presentational complexity already places us in a far better position for systematically exploring any differences in effect and design. The cohesive chain diagram generated directly from the annotated data for the segment of the video discussing the example slide is shown in Figure 6. It should be noted that this visualization “simply” gives a graphical rendering of the many cohesive links in the actual analysis, and so contains considerable information; this is generally the case for any complete cohesive analysis presented visually, even with monomodal verbal texts. The diagram is presented

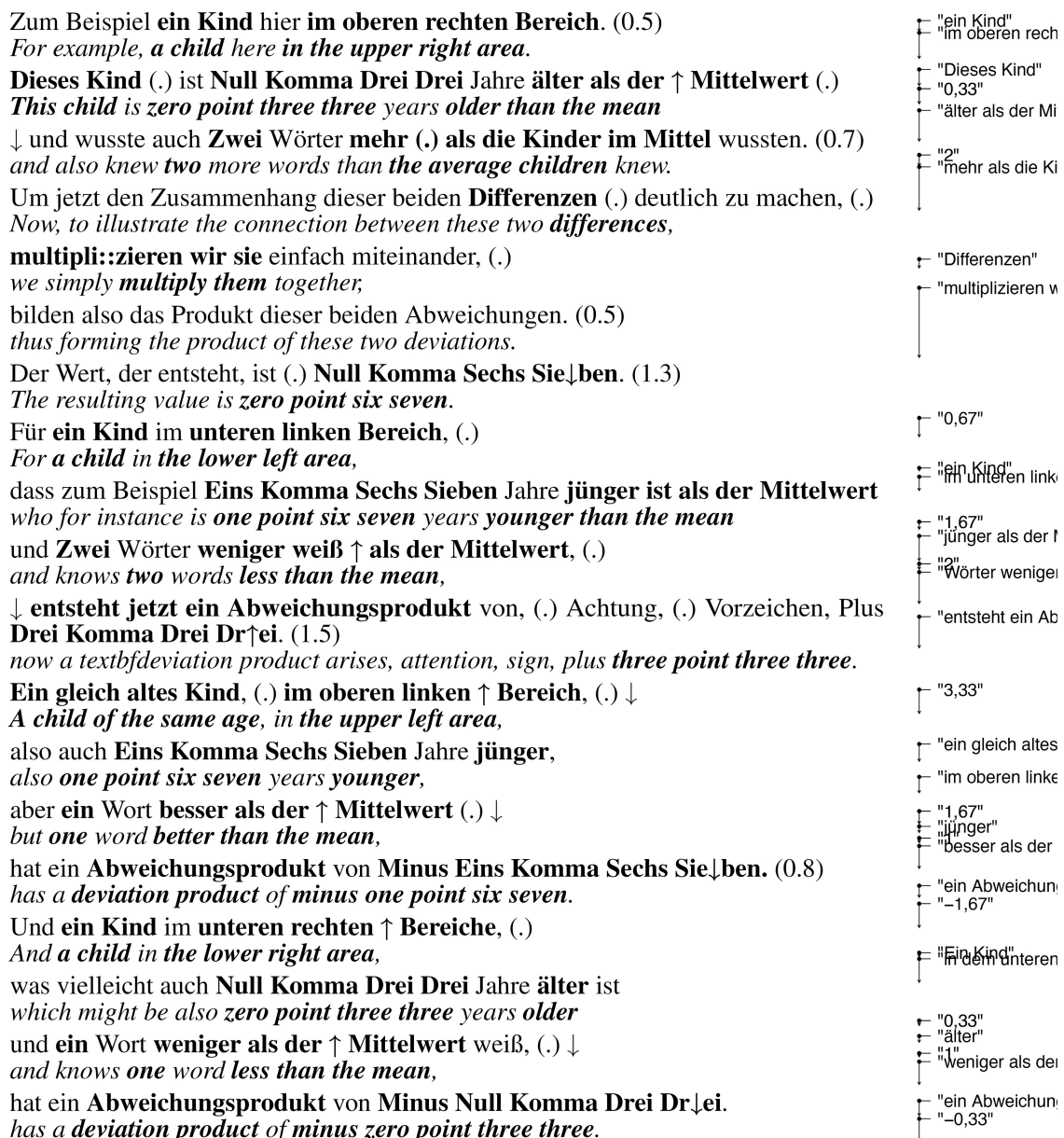


FIGURE 5

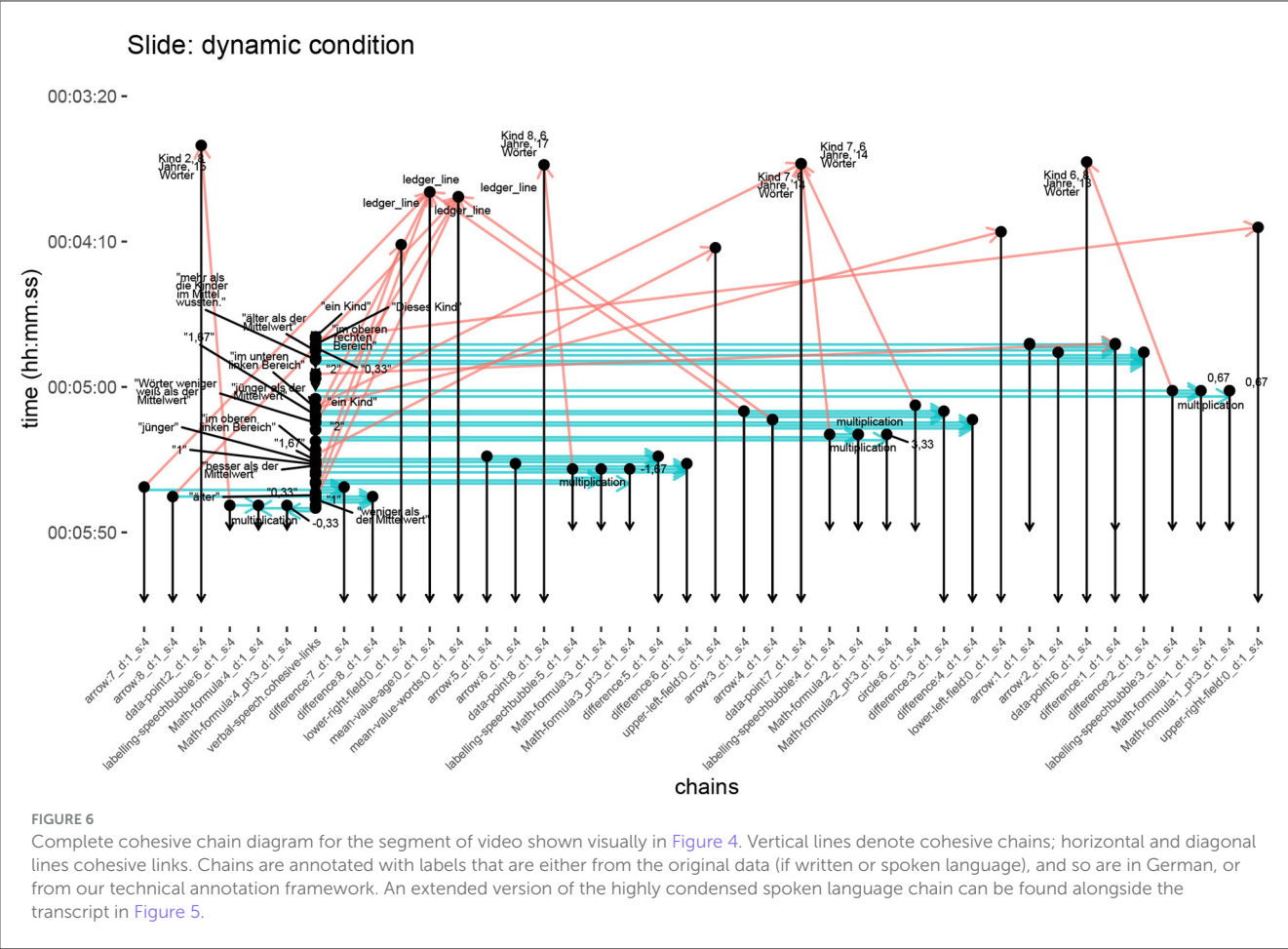
**Left:** Verbal transcript of the presentation accompanying discussion of the example slide in Jefferson notation, augmented additionally to show the phrases picked up in the cohesion analysis in bold. **Right:** the corresponding cohesive chain for the spoken language (with labels truncated right to save space and vertical extents corresponding to their actual temporal positioning: see below).

in full here to give a more realistic indication of the quantity of information being produced during analysis, although when working with particular areas from the overall data, it is generally more useful to extract smaller, more focused fragments of the overall data being discussed. This is the approach we take below when comparing the cohesion analysis with the eye-tracking data.

The conventions used in all of these diagrams remains the same, however. The vertically running arrows identify the various elements present in the visual field and the verbal speech, all with their respective temporal extents. Thus lines which are shorter

in height extend for shorter periods of the video. In the present case, many end at the same time, shown by the aligned lower arrow-heads at the bottom of the diagram, because at that point in the video the slide changes and the visual elements denoted are then no longer present. The very densely interconnected seventh tier from the left is the verbal speech tier, reflecting the fact that individual referring phrases will generally have a much shorter temporal extent than the visual elements being referred to and so it is challenging to present these together in a single static graphic. The fully extended chain can be seen on the right of Figure 5, where the individual phrases that can only be seen in part in the full





cohesion diagram are also identified<sup>1</sup>. In contrast, the individual multimodal references are spread out horizontally by virtue of their being contributed by different chains. The density of the verbal references shows well how, in this case, the speech chain serves as an organizational backbone for the presentation as a whole.

The horizontal or slanted arrows in the diagram show the phoricity relations that hold across the different chains according to the co-reference information maintained in the annotations. These phoric relations are also distinguished as explained above according to whether they are cataphoric, anaphoric, or co-phoric on the basis of the temporal information present in the annotated data; in the current example, there are no cataphoric relations to be seen. Anaphoric relations (shown in red) consequently are those horizontal arrows running upwards on the page, showing the co-reference to be a “referring back” to an element that was already present. Co-phoric references (shown in blue) on the other hand are then the horizontal

1 We should note here that this “chain” differs from traditional verbal cohesive chains in that it groups all the spoken contributions together, thus instantiating cohesion *by mode*. This can also be expanded to track particular referents more finely, but we omit this for the purposes of the current discussion.

arrows, which means those co-references hold between two elements that appear synchronously at the same time in the video. As an example, the fourth tier from the left commencing just before 00:05:50 shows a component of a labeling speech bubble (labeling-speechbubble:6\_d:1\_s:4) referring back to the third tier from the left, depicting a data point (data-point:2\_d:1\_s:4). The co-reference is then indicated by an arrow slanting upwards because the time of reference follows the time of visual presentation of the referent. The same holds for all other links shown operating between chains, including the spoken verbal information.

Applying the visualizations offers a succinct overview of the fine-grained annotation data, although, as noted above, it is often more revealing to focus in on particular combinations of elements as we do in our discussion of the relation between the discourse structure and the eye-tracking data below. It should also be noted that the cohesion diagrams discussed in this paper already only show the co-referentiality information so as to avoid overloading the diagrams presented still further. It is equally possible to pick out any of the cohesive relations present in the annotation, such as connection information or color cohesion, and so on. These details are omitted for current purposes and are, in any case, better shown interactively.

## 4 Experimental study

In this section we present the eye-tracking data that will be relevant below for our consideration of their interactions with the cohesion analyses. As indicated above, this data was gathered in a previous experimental study, conducted in German, exploring the effects of certain controlled variations in presentation styles. In particular, the study explored relations between the visual presence or absence of a lecturer in videos together with potential interactions with whether the slides used in the videos included animated elements or not. The overall aim of this study was to examine how the experimental conditions might influence viewer attention and, subsequently, learning effects. For the purposes of the present paper, however, we focus specifically on the eye-tracking results gathered concerning the contrast between the static and dynamic slide presentation conditions for the single selected slide introduced above (Figure 4); for further information about the sample study as a whole (see Schmidt-Borcherding et al., in preparation).

Bringing the previous study results together with our current objectives of relating cohesive structures with eye tracking data, our basic hypothesis is that instances of insufficient cohesion signaling should have detectable effects on the gaze behavior. A prime example of such insufficiency is when the formal co-reference relations attempt to span too great a temporal distance and so fail to effectively bring together the mode ‘doing’ the co-reference work and the mode being co-referenced. Hence, it is our assumption that differences between the sets of eye-tracking data gained from the two experimental conditions might be correlated with corresponding differences in the discourse organization. We address this hypothesis directly in Section 5 below.

### 4.1 Materials and methods

The (sub-)sample relevant for the purpose of this paper consisted of 22 students of education sciences (mean age = 24.71 years; 17 female) who participated in the study as part of a course requirement. Students were asked to learn about covariance and correlation with a ten-minute educational video consisting of 15 presentation slides shown on a 15" laptop screen. Several versions of the video were prepared, created previously by Florian Schmidt-Borcherding for the purpose of earlier experiments focusing on coherence. The results concerning two of the prepared video versions are relevant here; these varied coherence in two ways. In version A, called the “high coherence” condition, individual elements of the presentation slides in the video (sequential text elements, diagrams, circles, arrows, color coding etc.) occur *dynamically* and *synchronously* with the verbal speech. In version B, called the “low coherence” condition, the compositional elements of an entire slide being presented appear under static visual development conditions, i.e. elements do not occur successively but concurrently “all at once,” and are consequently not synchronous with speech. The contents of the slides and the verbal explanation were the same in both conditions (cf. Figure 5). The slides filled the whole screen while verbal instructional explanations were audible, but without the speaker being shown.

Participants were randomly assigned to the two experimental conditions, with 11 participants in each condition. The eye movements of each participant when engaging with the videos in the two conditions were recorded by having the participants wear eye-tracking glasses while learning with the video. For this, we used a head-mounted eye tracking system (Tobii Pro Glasses 2) with a sampling rate of 50 Hz<sup>2</sup>. The eye tracking glasses recorded (a) the gazes of both eyes (i.e., binocular) and (b) the visual stimuli in front of the students eyes. Students were tested in single sessions in a windowless room.

### 4.2 Data preparation

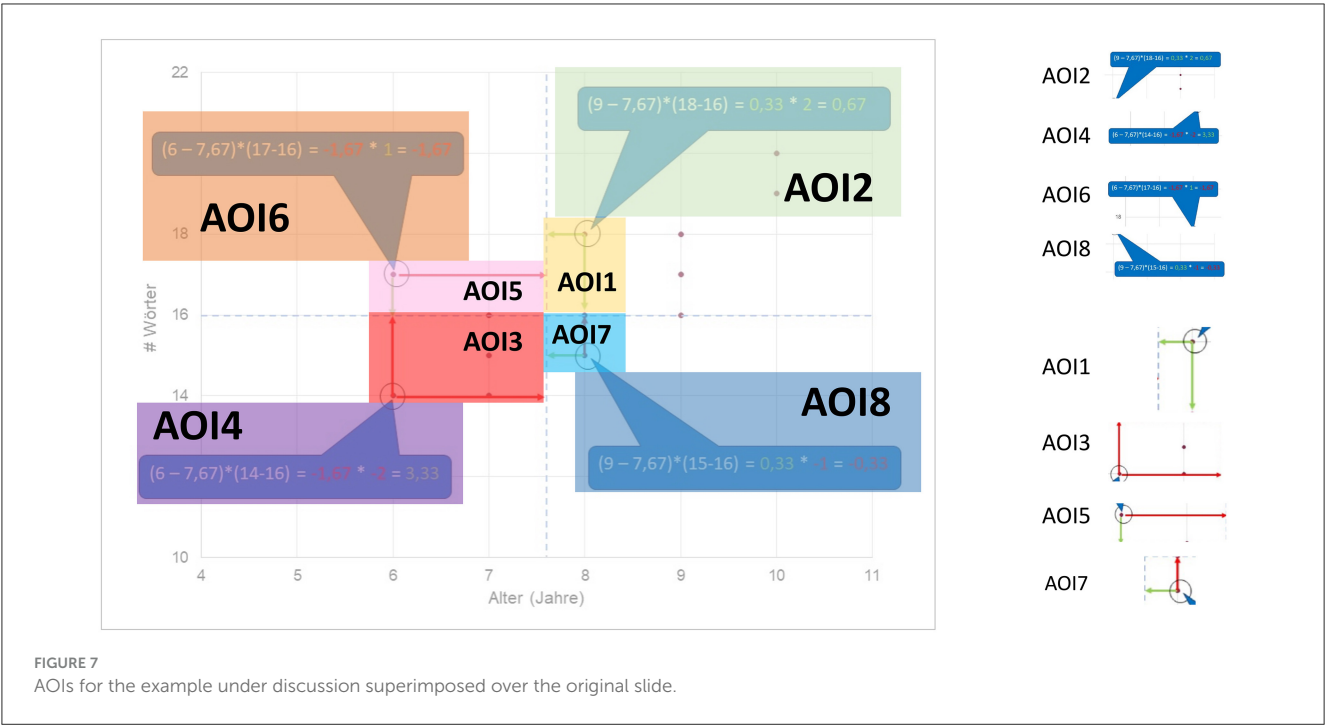
Calibration of the eye tracking system to the participants’ eyes failed in three cases. Hence, the further preparation and analysis of the eye tracking data discussed here is based on 19 valid data sets, ten in the high and nine in the low coherence conditions, respectively.

For the exploration reported here of relating both eye-tracking data and discourse analysis, we selected a particular sequence within the videos for detailed study, preparing the eye-tracking data accordingly. This segment concerns the presentation and explanation of the slide already shown in Figure 4 above. The interval of this video sequence extends from 04:42 min to 05:44 min in the videos as a whole, resulting in a duration of 62 seconds. When conducting eye-tracking experiments of this kind, it is generally beneficial to define particular spatial areas (*areas of interest*: AOIs) in the visual materials being used as stimuli so that gaze behavior can be contrasted specifically for those areas under differing viewing conditions. Consequently, we defined eight non-overlapping AOIs for our complex slide, positioned with respect to the slide as indicated in Figure 7. For ease of reference, these AOIs were numbered so as to follow the approximate ordering of the spoken discussion of those areas, although nothing hinges on this in the analysis. These AOIs themselves are then kept constant across the two experimental conditions of dynamic and static presentation so that any systematic differences found in gaze behavior can be attributed to variation in the conditions.

The sequence was also temporally divided into eight respective Times of Interest (TOI) to focus comparisons further. These TOIs are defined in broad correspondence to the AOIs: that is, a new TOI is defined as starting as soon as the first reference in the verbal speech is made to an element present in the corresponding AOI. Since the signaling function in the high coherence condition was applied by synchronizing dynamic elements with verbal speech, the TOIs also correspond with the onset of these elements.

As explained above, the segment at issue here is concerned specifically with presenting and explaining a graph depicting the divergences of four previously determined individual data points

<sup>2</sup> The choice of eye-tracker was made on the basis of the equipment available to us at the time the experiments were performed; in general, screen-based devices would have been more appropriate for the present study due to the increased tracking accuracy possible. We do not consider this to have had any deleterious effects for the results reported here, however, since finer resolution was not necessary.



from an overall set of 15 data points, representing children with varying ages and variable active vocabularies. The graph itself with its two axes of “age” and “vocabulary” had already been established earlier in the video’s discourse. The visual information focusing on the explanation of the divergences between data points and mean values was then added in one of two ways according to the two experimental conditions. In the dynamic version of the video the extra information was added successively and synchronously with the verbal speech; while in the static condition, the extra information was already present in the new slide. In the dynamic case, both the previously introduced parts of the graph and any newly appearing elements remained visible until the discussion of the divergences of each of the four data points had been concluded. As was seen in Figure 4, the divergences discussed are shown in the graph in terms of their relations to the mean values of the data set as determined through multiplication formulae presented in “speech balloons.” The data points and divergences then form the contents of the odd numbered AOIs shown in Figure 7; the mathematical formulae constitute the even numbered AOIs.

For the time intervals corresponding to this sequence, the raw eye tracking data were aggregated into fixations using the default value thresholds for fixation duration and recognition implemented in the Tobii Pro Lab Eye Tracking software. A fixation is defined as a time interval during which the eye gaze is maintained at a single location. Typically, those fixations alternate with so-called saccades, which are quick movements of the eye to the next fixation location<sup>3</sup>. During a saccade the eye is functionally blind. Hence, according to the eye-mind-hypothesis (cf. Just and Carpenter, 1980), the location of fixations can be interpreted as indicating the locus of

<sup>3</sup> More technically, the Tobii processing software considers fixations to be sequences of eye tracking data points where the velocity of eye movement lies below a given threshold, by default 30°/s; i.e., where there is not a saccade.

TABLE 1 Schematic matrix of the 64 AOI<sub>m</sub>/TOI<sub>n</sub> ( $m, n = 1, 2, \dots, 8$ ) variables aggregating single fixations across the analyzed video sequence.

	AOI <sub>1</sub>	AOI <sub>2</sub>	...	AOI <sub>8</sub>
TOI <sub>1</sub>	AOI <sub>1</sub> /TOI <sub>1</sub>	...	...	AOI <sub>8</sub> /TOI <sub>1</sub>
TOI <sub>2</sub>	...	AOI <sub>2</sub> /TOI <sub>2</sub>		
...	...		...	
TOI <sub>8</sub>	AOI <sub>1</sub> /TOI <sub>8</sub>			AOI <sub>8</sub> /TOI <sub>8</sub>
	Σ AOI <sub>1</sub>	Σ AOI <sub>2</sub>	...	Σ AOI <sub>8</sub>

Marginal sums for each AOI<sub>m</sub> for the whole length of the sequence (TOIs 1 to 8) are highlighted in green. The main diagonals of the matrix, containing the eight AOI<sub>m</sub>/TOI<sub>n</sub> (with  $m = n = 1, 2, \dots, 8$ ) combinations, is highlighted in orange.

visual attention, the duration of fixations can be interpreted as an indication of the amount of visual attention devoted to the locus of attention, and the sequence of fixations can be interpreted as the shift from one locus of visual attention to the next.

In a second step, we further aggregated fixations into the spatial and temporal dimensions of attention to the sequence. For each participant, we summed fixation durations on each AOI for each TOI, giving 8 (AOIs) × 8 (TOIs) = 64 variables representing the amount of visual attention devoted to a specific part of the slide during a specific time interval. The combination of AOIs and TOIs can be visualized in a matrix as shown in Table 1.

The present analysis aims to focus on the distribution of visual attention during the specific sequence. Absolute values of fixation durations may then be misleading for such analysis because of two possible measurement errors. First, even though the video has a fixed length, participants may vary individually in the absolute time they devoted to watching it. Second, even in the most reliable eye tracking measures there is still some data loss—that is, fixation durations do not necessarily sum precisely to the length of the

TABLE 2 Means (M) and standard deviations (SD) of relative fixation durations for high and low coherence conditions on AOIs: (A) as marginal sums for the whole length of the video sequence (left columns), and (B) for AOI<sub>n</sub>/TOI<sub>n</sub> ( $n = 1, 2, \dots, 8$ ) pairs.

	Marginal sums (length of sequence)				AOI <sub>n</sub> /TOI <sub>n</sub> ( $n = 1, 2, \dots, 8$ )			
	High coherence		Low coherence		High coherence		Low coherence	
	M	(SD)	M	(SD)	M	(SD)	M	(SD)
AOI 1	0.05	(0.06)	0.10	(0.11)	0.01	(0.03)	0.05	(0.06)
AOI 2	0.18	(0.16)	0.22	(0.13)	0.10	(0.08)	0.07	(0.10)
AOI 3	0.14	(0.08)	0.07	(0.04)	0.05	(0.06)	0.02	(0.02)
AOI 4	0.18	(0.10)	0.07	(0.07)	0.06	(0.05)	0.01	(0.02)
AOI 5	0.13	(0.07)	0.03	(0.03)	0.03	(0.05)	0.02	(0.02)
AOI 6	0.15	(0.11)	0.33	(0.23)	0.07	(0.07)	0.07	(0.05)
AOI 7	0.09	(0.09)	0.06	(0.06)	0.06	(0.09)	0.01	(0.01)
AOI 8	0.09	(0.10)	0.11	(0.08)	0.07	(0.09)	0.03	(0.04)

measurement sequence. To deal with these potential sources of variation, we calculated a relative attention distribution value for each of the 64 variables. This relative attention distribution was calculated by dividing the fixation duration of each specific AOI/TOI-combination (i.e.,  $\{AOI_m/TOI_n\}_{m,n=1,2,\dots,8}$ ) by the sum of all AOI/TOI-combinations for each participant.

### 4.3 Results

To statistically describe differences in viewing behavior between the two experimental conditions, we performed two Analyses of Variance (ANOVA). Both ANOVAs were conducted as a  $2 \times 8$ -factorial design with the between-subjects factor of coherence being high vs. low, and a within-subjects factor AOI referring to the AOIs 1 to 8, respectively.

In the first ANOVA, the dependent measure was the relative amount of visual attention devoted to each AOI over the whole length of the sequence. That is, for each AOI we summed up the relative fixation times from TOI 1 to TOI 8 for each participant. The descriptive statistics are shown in the left-hand group of columns in Table 2. The results revealed a significant main effect for AOI [ $F_{(7,11)} = 3.493$ ,  $p = .032$ ,  $\eta_p^2 = 0.69$ ], indicating that visual attention irrespective of the coherence condition is not evenly distributed across the AOIs of the sequence. This statistical result is illustrated in Figure 8A by the zigzagging line. The main effect for the between-subjects factor of coherence could not be calculated with the relative sum of AOIs over the whole sequence as the means in both conditions sum to 1. There is simply no descriptive difference between both conditions in the relative fixation times that could be further statistically qualified. Nevertheless, an effect of coherence is qualified by a significant interaction between AOIs and coherence [ $F_{(7,11)} = 4.643$ ,  $p = .012$ ,  $\eta_p^2 = 0.75$ ]. As can be seen in Figure 8B, visual attention appears more evenly distributed across the eight AOIs in the high compared to the low coherence condition. *Post-hoc t*-tests with the between-subjects factor coherence (high vs. low) for AOIs 1 to 8 respectively, revealed the differences in AOIs 4, 5, and 6 to be significant at a 0.05 level. After Bonferroni-correction ( $p = .05/8 = .006$ ), only the relative

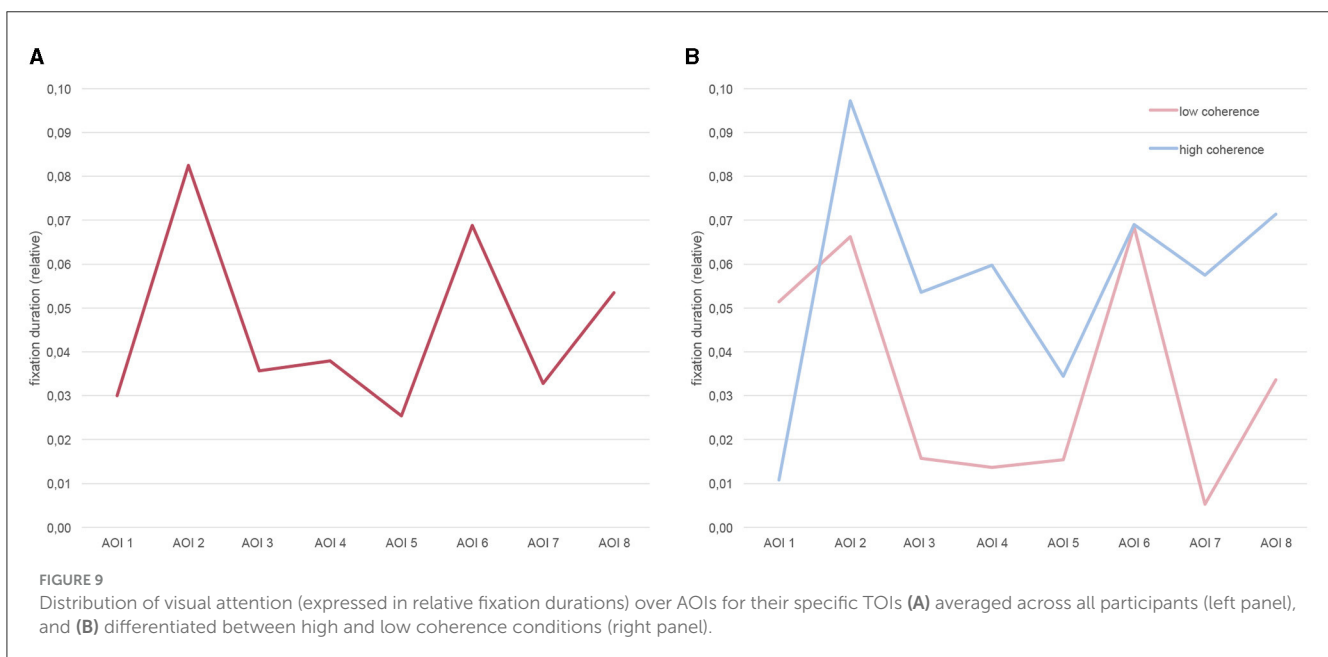
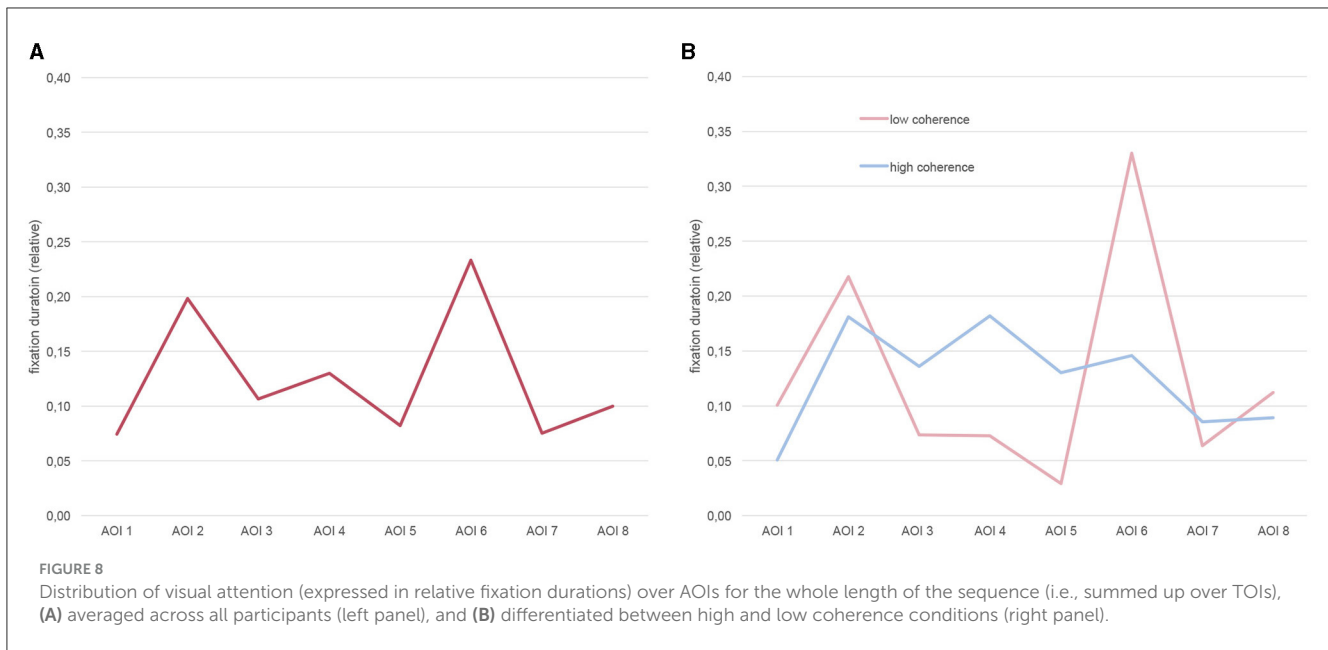
amount of visual attention paid to AOI 5 remained statistically different between both conditions, however.

In order to get a clearer picture of how attention distribution might be altered by coherence between the spoken language and the visual information we conducted a second ANOVA with just the eight AOI/TOI-pairs lying on the diagonal of the AOI/TOI-matrix (i.e.,  $\{AOI_n/TOI_n\}_{n=1-8}$ , cf. Table 1). The rationale for this analysis is the assumption that these pairs should better reflect the “cohesion path” of the video as grouped around both the spatial areas and the temporal intervals relevant for each AOI/TOI pair. The descriptive statistics for this are shown in the rightmost columns of Table 2. Results again revealed a significant main effect for AOI [ $F_{(7,11)} = 3.292$ ,  $p = .038$ ,  $\eta_p^2 = 0.68$ ], indicating that visual attention irrespective of the videos’ coherence is not evenly distributed across the AOIs of the coherence path. This statistical result is illustrated in Figure 9A, again by the zigzagging line. The ANOVA also revealed a main effect for coherence [ $F_{(1,17)} = 9.773$ ,  $p = .006$ ,  $\eta_p^2 = 0.37$ ] indicating that the amount of visual attention paid on the coherence path of a video varies with the coherence of the video. Indeed, while in the high coherence condition more than 45 percent of the measured visual attention was paid on the coherence path on average ( $SD = 15.53$ ), this attention sums up to only 27 percent in the low coherence condition on average ( $SD = 8.74$ ). This effect is illustrated in the right-hand panel of Figure 9. With the exception of AOI<sub>1</sub>/TOI<sub>1</sub>, the line for the high coherence condition is above the line for the low coherence condition. The interaction between AOIs and coherence failed statistical significance however, albeit only just [ $F_{(7,11)} = 2.903$ ,  $p = .056$ ].

### 4.4 Discussion

Taken together, the experimental study revealed that participants gaze behavior was affected by the two presentation conditions. Concerning the overall distribution of visual attention across the most relevant parts (AOIs) of the slides, signaling the relevance of these parts dynamically and synchronously with the verbal speech (i.e., the high coherence condition) led





to a more even distribution of attention compared to a static slide presentation lacking these signals (i.e., the low coherence condition). Although the  $\eta_p^2$ -value of this shift (i.e., the interaction) indicates this effect to be (very) large, we could hardly identify single AOIs to explain it, probably due to the effect working in both directions. Descriptively, four AOIs gain more attention (3, 4, 5, and 7) and four AOIs gain less attention (1, 2, 6, and 8) in the high compared to the low coherence condition. More remarkable from a descriptive perspective is that the divergence between the two conditions appears to be high in the middle AOIs (3–6) while the gaze behavior on AOIs 1, 2, 7, and 8 appears comparable. Intuitively, a growing divergence is reasonable since the time lag of appearance for the AOIs between high and low coherence conditions also grows from AOI 1 to AOI 8. However, the validity of this interpretation is called into question

by the similarly low attention devoted to AOIs 7 and 8 in both conditions.

Shifting the focus from overall attention distribution to a path-like measure revealed an even more differentiated picture of the participants' gaze behavior. First, participants in the high compared to the low coherence condition spent much more time on an AOI when it was first referred to (i.e., the AOI/TOI-pairs in the diagonal of the AOI/TOI-matrix, cf. Table 1). This shift indicates how much additional visual attention is pulled to these AOIs by signaling features. That is, the signaling causes the gaze behavior to more strongly follow the coherence path of the video, and, thus, presumably homogenizes the gaze behavior to be more similar between participants in the high coherence condition. While the conclusion “signaling draws attention” appears trivial at first glance, attention to an educational video is assumed to

serve the purpose of learning its content. That is, in order to understand the capability of signals to draw attention, we need to bring together actual attention allocation (i.e., the empirical gaze behavior) with the presumed functions of particular signals for the cohesive structure of a video. This is then what remains to be addressed by the formal multimodal description of the material that we now present.

## 5 Correlating the eye-tracking study and the cohesion study

So far in this paper we have provided two building blocks for approaching explanation videos empirically. First, we introduced a detailed annotation scheme for any explanation videos exhibiting multimodal complexity. Second, we showed differences in observed gaze behavior for video presentations contrasting with respect to their synchronization of visual information and accompanying spoken language. In this section we attempt to triangulate aspects of the discourse structure revealed by our annotation against the variation observed by the eye-tracking data.

This will serve several functions. First, it is necessary in general to provide empirical support for the kinds of distinctions shown in the discourse analysis; differences in discourse organization should correlate with differences in measurable behavioral factors among recipients. If this were not the case, then we have no basis beyond purely theoretical argument that the discourse analysis is actually capturing significant aspects of the objects analyzed. Second, and more specifically, if we can match formal properties of the discourse analysis with attention allocation, then we will be one step further toward being able to provide a systematic way of predicting to what extent particular video designs may help guide attention. By these means we may begin to isolate characterizations of signaling properties that are anchored both in fine-grained details of form and in predictions for reception effects. It must be noted, however, that the extent, if at all, that correlations can be found between a detailed cohesion analysis of multimodal text organization and the reception of texts so analyzed remains an open research question at the present time. Indeed, as we shall see, this is a complex undertaking that requires significant further work.

### 5.1 Cohesion analyses of the selected contrasting examples

As explained above, in order to organize the eye-tracking data for comparison across experimental conditions, it was useful to identify specific spatio-temporal segments for close attention. A similar range of considerations now needs to be applied to the cohesion analysis since the cohesion analysis of a segment of video provides only a snapshot of the relations holding between elements within that segment. That snapshot is “static” in the sense that all the cohesive relations constructed during the segment are recorded and made accessible for analysis.

To relate such “synoptic” analyses to the unfolding of a viewer’s understanding of a video, therefore, we need in addition to incorporate the theoretical construct of *logogenesis* (Halliday and

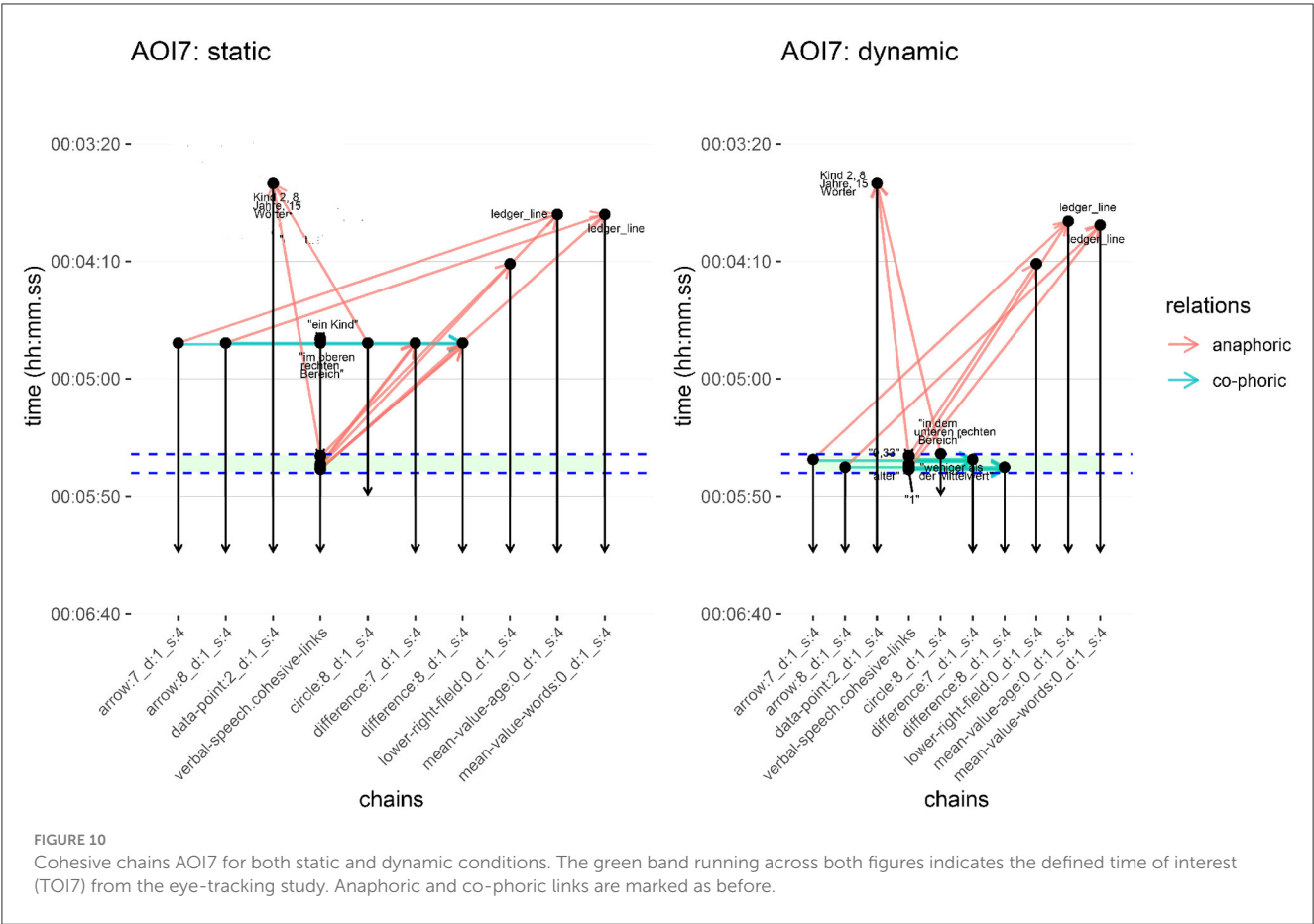
Matthiessen, 2013, 63), which sees texts articulating structures and relationships over time as they develop. For linear monomodal written text, this development is essentially one dimensional, although the structures and relationships constructed are generally more complex. In the multimodal audiovisual case, the situation is more complex still as the material can also include multiple simultaneous strands of development. These strands are what are captured in the overall cohesive analysis as illustrated above in Figure 6.

We can capture the dynamic unfolding of multimodal texts by augmenting the synoptic cohesion diagrams in two ways. First, we employ a notion of a textual “now” that moves successively downwards through the diagram as the text unfolds. Cohesive relationships that have not yet been made with respect to this “now” cannot have an influence on discourse construction and so are considered inaccessible for purposes of characterizing the point-by-point cohesive organization of the text. And second, we focus on just those portions of the cohesive analysis that are “present” with respect to the temporal “now.” This is particularly important for considering the consequences of the overall difference between the static and dynamic experimental conditions. Since the static condition does not allow any development of the contents of slides during the presentation, this corresponds to a restriction to the possibilities of the material of the medium: in short, the canvas (see above) is restricted so that it is not dynamic. As we shall see in a moment, this change in canvas means that the synoptic description of the cohesive relations applying may also change in quite specific ways.

The inclusion of these dynamic aspects allows us to draw a natural connection with the methods employed in the eye-tracking study. We first focus on portions of the overall cohesive analysis by selecting just those cohesive chains representing elements present visually within the defined AOIs. An AOI then corresponds to a subset of the cohesive chains in the cohesion diagram overall. We also define respective “now” intervals to correspond to the defined TOIs of the eye-tracking study. Combining these two aspects allows the temporal and the visuospatial restrictions to provide regions within the cohesion analysis that, on the one hand, may be used for calculating spatiotemporally restricted cohesion statistics of various kinds and, on the other hand, may be compared with the corresponding AOI/TOI figures provided by the eye-tracking data. We hypothesize that such cohesion statistics will differ according to whether the visual information is aligned with the spoken language or not.

The change in experimental condition also has consequences for the cohesive chain diagrams concerning the temporal intervals involved. Whereas in the static condition, some element would be present for the entire time that that visual is “on screen”—typically for the entire duration of a single slide, in the dynamic condition the same element may appear and disappear even within the duration of a single slide. This means that the corresponding cohesive chains have different temporal extents across the two conditions, resulting in connecting arrows of different lengths in the diagrams.

As a concrete example of this, we can contrast the diagrams constructed for a selected area of interest in both the static and dynamic video conditions. Figure 10 shows corresponding cohesion diagrams for AOI7 in the two conditions. As noted above, the basic multimodal “make-up” of odd and even numbered



AOIs respectively is consistent within each group in that the odd numbered AOIs refer to the data points and the even numbered AOIs refer to the mathematical call-outs. In this sense, AOI7 serves as an exemplar for all odd-numbered AOIs. However, AOI7 is also a region that exhibited a considerable difference between the static and dynamic conditions in the eye-tracking data (cf. Figure 9) and so is an interesting case to consider. Diagrams for all the AOIs from the study in both conditions are given in the [Supplementary material](#).

Contrasting the two diagrams for AOI7 reveals two prominent points of difference. First, among the visual elements forming cohesive chains, there are temporal disparities across the static and dynamic conditions concerning both the occurrence of those visual elements and, consequently, their establishment of references. And second, the distribution between anaphoric and co-phoric relations appears quite different. Under the static condition, the cohesive references are primarily of an anaphoric nature (giving 11 such relations in total), while co-phoric references are notably sparse (2 in total), occurring around the 4:50 minute mark; conversely, in the dynamic condition this distribution reverses giving a more balanced distribution between co-phoric (6) and anaphoric references (7). Thus, in the static condition, visual cohesive chain elements appear “earlier” and so establish references to other elements in different cohesive chains earlier as well. In the dynamic condition, the same visual elements occur and build references “later”—in this case almost exclusively within the time of interest defined for the AOI as indicated in the figure. The verbal elements

TABLE 3 Counts of anaphoric and co-phoric references per AOI and per condition (dynamic vs. static).

AOI#	Dynamic		Static	
	Anaphoric	Co-phoric	Anaphoric	Co-phoric
AOI1	8	6	8	6
AOI2	2	4	5	1
AOI3	7	6	11	2
AOI4	1	3	3	1
AOI5	7	6	11	2
AOI6	1	3	3	1
AOI7	7	6	11	2
AOI8	1	3	3	1

offer an exception to this general contrast, however, since they engage in phoric relations at the same time in both conditions. Their temporal information therefore remains the same, although the specific types of phoric relations holding are determined by the relative temporal positions of the elements being related and so vary.

This pattern of difference can be observed across all the AOIs, as can be seen visually in the graphs in the [Supplementary material](#). But we can also capture this quantitatively by considering the overall counts of the different phoricity types between the static

and dynamic conditions and across the defined AOIs. These are tabulated in [Table 3](#). We see here that the number of anaphoric and co-phoric references under the static condition differ greatly compared to those in the dynamic version in general. This can be explained due to the fact that in the static case, visual elements do not occur successively but all at the same time and so appear earlier in the discourse compared to the dynamic condition where, again, elements occur synchronously to verbal speech. The temporal development of the verbal speech itself in both versions remains the same. Taking these points into consideration, consequently, there are more anaphoric references present in the video under static conditions and far fewer co-phoric ones, whereas under dynamic condition this distribution is more leveled.

Another way of bringing out the differences in cohesion analysis across the two experimental conditions is to compare the *proportions* of the distinct types of phoricity relations. For this, we scale the absolute count of phoricity relations, anaphoric or co-phoric in the present case, by the number of cohesive chains in the focused area of interest overall. The reason for this is to avoid over-scoring particular regions simply because they have more elements being related. This is a loose correlate of scaling done for the eye-tracking data concerning the absolute sizes of the areas of interest. Whereas a larger area might be expected to receive more fixations simply by chance, here we might expect there to be more cohesive relations simply because of more elements being present. Graphs of the scaled results, separating out anaphoric and co-phoric relations across both the experimental conditions and the AOIs are shown in [Figure 11](#). Here the difference in behavior is very clear. In the static condition the proportion of co-phoric references dramatically decreases after AOI1; since the visual elements are, by and large, already present, most relations are necessarily anaphoric, although there is systematic variation between the odd and even AOIs, again reflecting their distinct compositions as set out above. In the dynamic case, there is a far more even use of co-phoric and anaphoric cohesive links. Considering just the phoric relation totals as shown on the right of the figure, the difference in use across conditions is highly significant ( $\chi^2 = 12.042$ ,  $df = 1$ ,  $p = .0005$ ).

The graphs also suggest that there are marked differences between the odd and even AOIs. This is suggestively similar to the differences in fixation times observed in the eye-tracking data above. As noted there, the even and odd AOIs are rather different in their multimodal composition (cf. [Figure 7](#)). This appears to be reflected in their cohesive properties as well, although the situation quantitatively is less clearcut. Grouping the odd and even AOIs and comparing those groups' total phoricity counts within conditions and across anaphoric and co-phoric references shows a mixed picture. In the dynamic case, differences in the counts for anaphoric and co-phoric relations fall just short of significance at the 95% level (Fisher's exact test,  $p = .0596$ ). For the static case, no significant difference in raw counts is found at all (Fisher's exact test,  $p = 1$ ). Similarly, looking instead at the *proportions* of anaphoric and co-phoric relations with respect to the total number of available elements across the even and odd AOIs in the two conditions shows an identical pattern: the proportion of anaphoric references increases from dynamic to static, and the proportion of co-phoric references decreases from dynamic to static. Thus, although the counts and the proportions are different for the odd and even AOIs, the pattern of change remains the same and so we will not consider

these differences in AOIs further below. More data exhibiting these and other differences in multimodal composition might well show differences, but from the data at hand we can make few predictions.

## 5.2 The relation to the eye-tracking results

The analysis so far demonstrates that there are substantial differences between the two conditions in terms of their respective cohesion analyses. The contribution of the various AOIs appears of less significance. It remains to be seen, however, whether any of these differences stand in any specific relation to the differences found in the eye-tracking study.

In order to explore this question, the cohesive relation data was augmented further with time-dependent information to reflect more accurately the unfolding nature of the “text” and its logogenesis. The starting and ending points of each cohesive relation present in an AOI analysis were classified with respect to their temporal position relative to the specified time of interest of that AOI. The temporal relations adopted were based on the standard relations from [Allen \(1983\)](#) interval calculus: i.e., intervals may overlap (extending before, after, or both), be entirely contained one within the other, or be disjoint. Following this step, each cohesive relation was annotated additionally according to its phoricity status, the respective time intervals of the elements between which the cohesive relation holds, and the temporal ordering relation of these two intervals with respect to the relevant TOI.

There are then several possibilities for evaluating the data further. To begin, we can again examine the cohesion data “internally” to see if there are other relationships among the calculated features to be brought out. For current purposes this was done by creating generalized linear models to see if selected dependent variables can be “predicted” from other variables in the data. An informal description of this process for multimodal data is given by [Bateman and Hiippala \(2021\)](#); technical details of the technique are, for example, given by [Baayen \(2008\)](#). Following this method, we first examined whether any combinations of the just described annotations added for each cohesive link would function as effective predictors of the experimental condition. That is, we see to what extent the annotation properties group differently according to whether they are drawn from the static or the dynamic condition. On the basis of the visualizations of the cohesive relations for the AOIs given above and in the [Supplementary material](#), one would expect this simply because the configurations look very different.

The regression model produced in this case indeed shows that there is indeed a significant contribution to the prediction of the condition as being either “static” or “dynamic” made by the starting interval of the cohesive relation when it is positioned either inside ( $p = .04$ ) or overlapping *after* ( $p = .0002$ ) the respective time of interest (see the [Supplementary material](#) for the full model). There is also a significant contribution for the interval to which the cohesive relation is referring when that interval is overlapping *after* as well ( $p = .04$ ). A small contribution ( $p = .07$ ) is also made by AOI1, which stands out from the other AOIs as already indicated in several of the graphs and counts above. These results



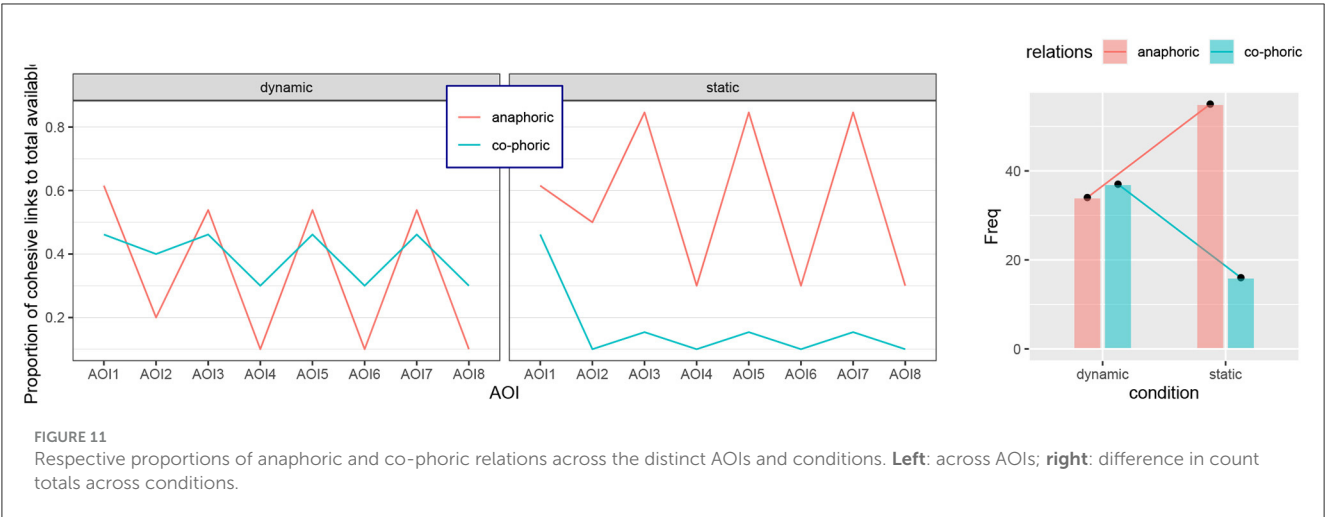


FIGURE 11  
Respective proportions of anaphoric and co-phoric relations across the distinct AOIs and conditions. **Left:** across AOIs; **right:** difference in count totals across conditions.

TABLE 4 Mixed effects model for predicting gaze duration on the basis of phoricity.

Formula: value ~ anaphoric + "co-phoric" + (1   AOI)					
Fixed effects:					
	Estimate	Std. Error	df	t value	Pr(> t )
(Intercept)	0.063778	0.013375	6.253648	4.769	0.00277**
Anaphoric	-0.005689	0.001522	9.201333	-3.738	0.00446**
Co-phoric	0.003951	0.002327	12.566915	1.698	0.11409
---					
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

The probability of the named variable's contribution to the prediction including zero is shown in the final column. The Estimate column gives the coefficients of the named variables, i.e., just how much they influence the prediction. The other values in the table give a sense of the variability and significance of that influence. The full table is available in the [Supplementary material](#).

are consequently more or less direct corroborations of the visual impressions from the contrasting graphs.

More relevant for our purposes of triangulation is then the relationship between the cohesion configurations and the eye-tracking data. For this, we take the relative gaze duration figures for the two conditions and the various AOIs as given in the right-hand columns of [Table 2](#) above. We then seek to see if these values can be "predicted" by some combination of the annotated cohesive features. To the extent that it is possible to predict durations on this basis, we would have shown that the cohesion analysis offers a proxy for at least some of the behavioral properties that can be measured in reception studies. To establish this prediction, we construct generalized linear models as before, exploring several possibilities.

For our first model we took the same annotated cohesion data as above but used the *relative duration times* as the values to be predicted rather than the experimental conditions. Since the distinct AOIs had not been found to show any particular individual influences before, we now took them as a random effect to produce a mixed effect model. Random effects in a generalized linear model are used to capture variables whose values are not expected to contribute systematically to a prediction, instead contributing "noise" potentially reducing the efficacy of the model as a whole ([Baayen et al., 2008](#)). The results were then identical to the previous internally generated model: here again the TOI-dependent relations of "overlapping after" and "inside" showed themselves to make statistically significant contributions to a prediction of relative duration. This is consequently interesting as

a first triangulation step, suggesting that there are indeed systematic correlations to explore.

To test this further, we next considered a variant of the previous model that instead grouped all of the results for the distinct AOIs together rather than keeping them separate. A model of this kind might plausibly reflect our experimental situation more accurately since we hypothesize that it will be overall cohesive properties of an AOI that play a deciding role and not individual cohesion relations. The data here was therefore aggregated by conditions and AOIs so that counts for each of the phoricity relations were available; these were similar to the counts given in [Table 3](#) but excluded any references to, or from, relations fully outside of the time of interest. The results here need to be treated with some caution, however, as we then only had 16 individual data points (i.e., 2 conditions across 8 AOIs). Moreover, the relative temporal information appeared to mask, or outweigh, the phoricity information as only the former appeared (again) to have a statistically significant effect. Consequently, to focus on these relations more particularly, we constructed a final mixed effects model that only related phoricity relations with the relative gaze durations for the conditions and AOIs, treating AOIs as random effects as before. The results of this model are summarized in [Table 4](#) (the full table is given in the [Supplementary material](#)).

Here we can see that there is, indeed, an apparently (highly) statistically significant contribution from anaphoric references ( $p = .004$ ), although their effect on the corresponding relative gaze duration remains small. Nevertheless, we can take this as at least preliminary supporting evidence that a description of the

development of the video in terms of multimodal cohesion may leave measurable effects on properties such as gaze duration. It is interesting in the current case that the experimental condition did not make a significant contribution when added to the model. This may fit well with the “mixed” nature of the experimental stimuli. For example, as we have discussed above, it is not the case that all AOIs behave differently in the two conditions: the material presented concerning AOI1 appears at the very beginning of the video segment analyzed and is consequently uniform across both conditions. This might restrict the ability of the model to distinguish cohesion configurations on the basis of the condition, but the effect of anaphoricity on relative gaze duration remains. Here we need to move to more corpus-oriented evaluations applying the same techniques as set out here but on a larger scale.

Considering the results overall, however, we can now attempt to make some preliminary hypotheses concerning how the cohesive analysis and the eye-tracking data may be brought into closer alignment. As discussed in the discussion of the eye-tracking results, the “high coherence,” or dynamic, condition appears to raise the allocation of attention to the AOIs concerned considerably; this was evident in the difference shown between attention on the “diagonal” components of Table 1 in the two conditions. When examining the cohesion analysis of the two conditions as summarized in Figure 11, there also appears to be a considerable difference between the conditions and, in particular, with respect to the way in which the anaphoric and co-phoric contributions relate.

In short, in the dynamic condition, there is a considerable overlap among the number of anaphoric and co-phoric contributions active across the AOIs. In contrast, in the static condition, these respective contributions quickly separate, leaving the large majority of cohesive links to be filled in anaphorically. This suggests the hypothesis that maintaining co-phoric relations may well increase the likelihood of attention being maintained and could even serve as a beneficial scaffolding device encouraging information integration. This is quite plausible and corresponds well to the general notions of signaling and cueing described above; here it is additionally significant, however, that we have begun to show how such results may be generated by empirical triangulation. Moreover, in terms of potential refinements for eye-tracking studies, this could well be explored further by paying particular attention to, for example, integrative saccades within AOIs across the contrasting conditions as suggested for quite different media by Holsanova et al. (2008).

## 6 Discussion and explorations

There are clearly still considerable issues of both theoretical and practical import to consider in the relationship between multimodal cohesion patterns and their potential input to the comprehension process. Something of the nature of this gap can be shown by explicitly contrasting the overall metrics obtained from cohesion analysis for the AOIs in the two conditions with the relative gaze duration graphs in Figure 9 above. As would be expected, the relative gaze duration figures show much more variation than that derived purely from the cohesion analysis. These can be compared directly by examining a combined measure of the contribution of cohesion shown in Figure 12. The values in

this figure are derived from the respective anaphoric and co-phoric proportions in a manner that attributes higher “scores” when the differences between anaphoric and co-phoric proportions are small, and lower “scores” when the phoricity relations are further apart. The circled points show the AOIs where this metric is equal across conditions. While some of these correspond approximately with the relative gaze durations, there are many cases which do not. We see, for example, that AOI6 scores equally across conditions, whereas AOI7 scores maximally differently. Although this aligns well with Figure 9, the equal scores of AOI1, AOI4 and AOI8 clearly do not align and so the model needs further refinement.

Many issues concerning how we might progressively bring the results closer together relate to aspects of logogenesis, i.e., the way in which we can formally characterize how a multimodal text is developing. There are a number of places where this may be expected to have significant consequences for attention allocation, and so dealing with each of these may improve the match between behavioral measures such as eye-tracking on the one hand, and the formal discourse analysis on the other. This in fact offers a research agenda with particular concrete steps for future investigation.

An illustration of the crucial role of logogenesis and explicit consideration of the “unfolding” of the text is offered by Figure 13. This cohesion diagram shows the cohesive links between chains for the static case of our area of interest AOI3. Here we can clearly see the potentially problematic phenomenon discussed above where visual material is introduced but only referred to verbally much later in the video’s development. Thus, on the left-hand side of the cohesion diagram we see the chain constructed by the verbal language containing a densely packed sequence of references to various visual aspects of the slide being presented as also seen in several of our diagrams above. These connections appear as anaphoric (red) links back to the respective chains of those referents. However, we also see two co-phoric relations (blue) among the visual elements occurring between 4:40 and 4:50 min. These ties are established by two arrows shown on the presentation slide (cf. Figure 4) that function as visual depictions of distances between values in the graph.

In our present scheme, establishing the status of ties as either anaphoric or co-phoric relies upon the strict temporal relations holding between the temporal intervals of the elements involved. This is evident in the diagram since the verbal references clearly follow the appearance of the referenced visual elements as shown by the earlier beginning of the corresponding vertical chains, and so are classified as anaphoric, whereas the co-phoric ties appear because co-referential visual elements appeared at the same time. While this is formally correct, such information may be dealt with differently by viewers because links may *only become relevant* when corresponding verbal references are made: up until that point, the information is visually present but, quite possibly, unattended to. This means that certain relations may be *formally* anaphoric and co-phoric as described, but may in reception function co-phorically when triggered within the time of interest indicated because this is when the corresponding verbal references occur.

Consequently, on the one hand, there may be conditions under which a visual element that is already present (and hence formally anaphoric) may function analogously to a co-phoric relationship when referenced verbally. Nevertheless, on the other hand, the fact

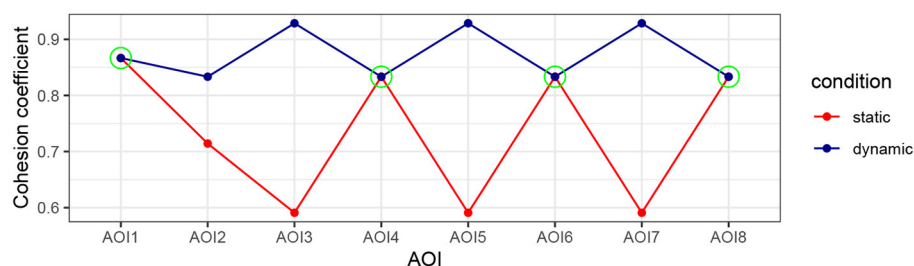


FIGURE 12  
Aggregated cohesion metric for the two conditions across AOIs.

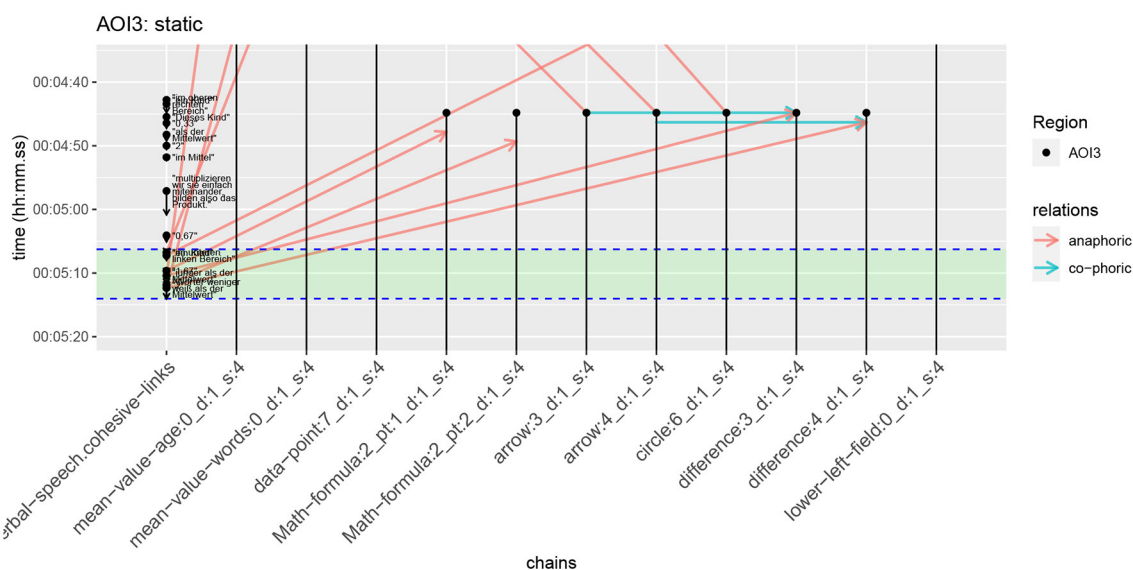


FIGURE 13  
Cohesive chains present in the static version of AOI3. The defined time of interest is marked by the green band; cohesive references are the horizontal or diagonal lines connecting vertical cohesive chains as above.

that a visual element appears just when it is referenced is also likely to exhibit a distinct and additional signaling effect of its own. That is: material that has been in the visual field for some time (as holds in the present case) and cases where the information freshly appears (which occurs in cases discussed above where the relevant cohesive chains begin within the time of interest) may need to be treated differently. Characterizing the consequences of these differences in the formal discourse analysis is then a clear challenge for the future and is consequently now placed prominently on our research agenda.

We begun in the previous section to treat such cases by additionally incorporating temporal relations calculated with respect to the specified TOI. This means that the vertical chains involved in the co-phoric relationships shown in the figure are annotated temporally as standing in a “contains” relationship because of the overlap seen between their temporal extents and the TOI. But we were not able to derive further significant results concerning the effects of such temporal relationships. This may be due to the extreme nature of the experimental contrasts adopted. The static condition often shows, as visible in the figure, no

co-phoricity, whereas the dynamic condition is already highly synchronous. In designs where this degree of synchronicity is not achieved, it may well be the case that we would begin to see more effects of potentially conflicting temporal information.

We will need to engage with the effects of such configurations more deeply. We might usefully consider, for example, the varying conditions under which the formally distant co-phoric relations visible in Figure 13 may be “brought into” the green zone of temporal interest by verbal (and other) signaling. If, for example, there were many potential referents in the visual field and the verbal signal was insufficiently precise to discriminate effectively, then one would expect that the difference between an element already being visually present and appearing temporally synchronized would have greater consequences. Such differences have also been studied in the area of multimodal referring expressions (e.g., van der Sluis and Krahmer, 2007), which could well provide useful additional input. This is also where we would place further signaling strategies such as spoken prominence, deictic gestures, and similar.

It is also likely that it is not only the bare presence of an anaphoric relationship that plays a role, since such relationships

may stretch over very different temporal intervals, corresponding to more straightforward “distance” in linear monomodal written text. When the differences in such temporal intervals become large, as is often the case in static presentation slides, corresponding references may fail to serve as effective guides of attention. In contrast, anaphoric references with small temporal differences between participating elements may then merge functionally with the effects of co-phoric references. Given this hypothesis that more “distant” anaphoric references then might not effectively guide attention, whereas “close” anaphoric as well as co-phoric ones might, our results could be further refined by employing restrictions responsive to these temporal differences. As there are, as discussed above, greater temporal differences between two elements engaging in a phoric relation under the static condition, this would motivate the hypothesis that for most of the anaphoric references in the static case, cohesion has not been so successfully signaled and thus attention was not being guided well.

Just how “forcibly” potential cohesive relations may be brought into the relevant temporal interval may then depend on a range of “signaling” or cueing factors that can now be studied more systematically for their effects. Moreover, as mentioned in Section 3.2 above, we have focused for current purposes specifically on multimodal cohesive relationships based on co-referentiality, but there are several further kinds of relationships that it would be natural to add, such as verbal stress and intonation on certain cue words, or cohesion based on form (e.g., color), and linking cohesion, such as causal and other types of connectives. All of these would be hypothesized to have effects suggesting that certain textual elements stand in specific relations to one another and so extend the texture of our discourse representation. These should all in the future be brought together for a more complete account of discourse signaling within a single integrated framework, for which the scheme defined here offers a robust foundation.

## 7 Conclusions and future work

In this paper, we have extended a previously introduced systematic, fine-grained annotation framework (Bateman et al., 2021) designed for the purpose of generating cohesion structures in explanation videos and explored whether differences in cohesive organization could be related to gaze data. For this, we drew on results from an earlier eye-tracking study which measured gaze behavior among a set of subjects for specific areas of interest of an explanation video. Comparing those eye-tracking results with the cohesive discourse structures of those AOIs supported the notion that fixation duration on areas that were referenced in the discourse in a synchronous manner showed differences characterizable in terms of phoricity relations. However, this hypothesis necessitates considerable further quantitative evaluation as well as extensions concerning the circumstances and variables to be drawn from the discourse analysis. Several directions for such extensions were outlined in the discussion in Section 6.

Our preliminary findings, upon which future work can build, suggest that co-references between two textual elements that are established more or less synchronously to the emergence of those elements will be more in line with established signaling principles. This means that textual elements referencing each other should

appear within a relatively similar timeframe in the discourse of the video (i.e., be co-phoric references) and be discursively coherent as well. This offers potential support most directly for Mayer’s principle of temporal contiguity (Mayer, 2009), but goes considerably further in tying our findings to specific identifiable elements in the overall multimodal presentation, rather than to notions of “text” and “image” as units. This makes it possible to pursue more fine-grained extensions of signaling accounts by examining more closely differences brought about by both the form of elements standing in cohesive relations and their precise temporal relationships, as set out in Section 6. This should allow us in subsequent research to scrutinize just those discourse structures that offer the most effective signaling possibilities at arbitrarily fine scales as might be needed for individual presentations..

Although we have outlined discourse structures that can serve as a scaffold for placing previous proposals for signaling principles in order to probe them further empirically, the present study has only focused on two possible and very distinct presentation styles for explanation videos. Empirical data involving a wider range of “mixtures” between the extreme cases of synchrony and non-synchrony of spoken language and visuals need now also to be considered systematically. This may then assist in reducing the “gap” observed between the current predictions that we can make on the basis of a thin slice of cohesion analysis and the variations found in the eye-tracking data by adding the kinds of effects and refinements to the model discussed in Section 6. Taken together, these points serve to define a set of clear research goals offering potentially beneficial results both for the practical task of characterizing explanation video design in a manner supportive of predictions concerning attention and effect and for the theoretical goal of improving the nature of multimodal discourse analyses.

In the future, therefore, through triangulating pedagogic, linguistic and multimodal theories for methodological purposes, we aim to establish more robust foundational frameworks capable of serving as a meta-language for annotations of empirically observable audiovisual linguistic phenomena relevant for theoretical learning principles as well. When applied to larger corpora, such a meta-language may then be standardized for broader quantitative research designs. Given the increasing prevalence of audiovisual learning materials, which present an intricate and challenging terrain for empirical research concerning their facilitation of positive learning outcomes, this undertaking is certain to become ever more important.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

## Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. All subjects gave informed consent to participate in the study. The study was conducted in accordance



with the code of ethics of the German Educational Research Association (DGfE) as well as APA ethical standards. All data were collected and analyzed anonymously.

## Author contributions

LT: Formal analysis, Methodology, Writing – original draft, Writing – review & editing, Conceptualization. FS-B: Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing, Conceptualization. JB: Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing, Conceptualization.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships.

## References

- Acartürk, C., Taboada, M., and Habel, C. (2014). Cohesion in multimodal documents: effects of cross-referencing. *Inf. Des. J.* 20, 98–110. doi: 10.1075/idx.20.2.02aca
- Ainsworth, S. (2008). “The educational value of multiple representations when learning complex scientific concepts,” in *Visualization: Theory and Practice in Science Education*, eds. J. K. Gilbert, M. Reiner, M. Nakhleh (Cham: Springer), 191–208. doi: 10.1007/978-1-4020-5267-5\_9
- Ainsworth, S. (2021). “The multiple representations principle in multimedia learning,” in *The Cambridge Handbook of Multimedia Learning* (Cambridge, UK: Cambridge University Press), 158–170. doi: 10.1017/9781108894333.016
- Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Commun. ACM* 26, 832–843. doi: 10.1145/182.358434
- Alpizar, D., Adesope, O. O., and Wong, R. M. (2020). A meta-analysis of signaling principle in multimedia learning environments. *Educ. Technol. Res. Dev.* 68, 2095–2119. doi: 10.1007/s11423-020-09748-7
- Ayres, P., and Sweller, J. (2021). “The split-attention principle in multimedia learning,” in *The Cambridge handbook of multimedia learning*, ed. R. E. Mayer (Cambridge, UK: Cambridge University Press), 199–211. doi: 10.1017/9781108894333.020
- Baayen, R. H. (2008). *Analyzing Linguistic Data. A practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511801686
- Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effect modeling with cross random effects for subjects and items. *J. Memory Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005
- Bateman, J. A. (2013). “Multimodal corpus-based approaches,” in *The Encyclopedia of Applied Linguistics*, ed. C. A. Chapelle (Hoboken, NJ, USA: Blackwell Publishing Ltd.), 3983–3991. doi: 10.1002/9781405198431.wbeal0812
- Bateman, J. A. (2022). Growing theory for practice: empirical multimodality beyond the case study. *Multimodal Commun.* 11, 63–74. doi: 10.1515/mc-2021-0006
- Bateman, J. A., and Hiippala, T. (2021). “From data to patterns: on the role of models in empirical multimodality research,” in *Empirical Multimodality Research: Methods, Evaluations, Implications*, eds. J. Pflaeging, J. Wildfeuer, and J. A. Bateman (Berlin: de Gruyter), 65–90. doi: 10.1515/9783110725001-003
- Bateman, J. A., and Schmidt-Borcherding, F. (2018). The communicative effectiveness of education videos: towards an empirically-motivated multimodal account. *Multimodal Technol. Inter.* 2:59. doi: 10.3390/mti2030059
- Bateman, J. A., Thiele, L., and Hande, A. (2021). Explanation videos unravelled: breaking the waves. *J. Pragmatics* 175, 112–128. doi: 10.1016/j.pragma.2020.12.009
- Bateman, J. A., Wildfeuer, J., and Hiippala, T. (2017). *Multimodality-Foundations, Research and Analysis. A Problem-Oriented Introduction*. Berlin: de Gruyter Mouton. doi: 10.1515/9783110479898
- Bétrancourt, M., and Benetos, K. (2018). Why and when does instructional video facilitate learning? A commentary to the special issue “developments and trends in learning with instructional video.” *Comput. Hum. Behav.* 89, 471–475. doi: 10.1016/j.chb.2018.08.035
- Breslyn, W., and Green, A. (2022). Learning science with youtube videos and the impacts of Covid-19. *Discipl. Interdisc. Sci. Educ. Res.* 4:13. doi: 10.1186/s43031-022-00051-4
- Bucher, H.-J., Krieg, M., and Niemann, P. (2010). “Die wissenschaftliche Präsentation als multimediale Kommunikationsform,” in *Neue Medien-neue Formate. Ausdifferenzierung und Konvergenz in der Medienkommunikation, number 10 in Interaktiva. Schriftenreihe des Zentrums für Medien und Interaktivität (ZMI)*, Gießen, eds. H.-J. Bucher, T. Gloning, and K. Lehnert (Frankfurt and New York: Campus Verlag), 381–412.
- Bucher, H.-J., and Niemann, P. (2012). Visualizing science: the reception of PowerPoint presentations. *Visual Commun.* 11, 283–306. doi: 10.1177/1470357212446409
- Castro-Alonso, J. C., and Sweller, J. (2021). “The modality principle in multimedia learning,” in *The Cambridge handbook of multimedia learning*, ed. R. E. Mayer (Cambridge, UK: Cambridge University Press), 261–267. doi: 10.1017/9781108894333.026
- Djonov, E., and van Leeuwen, T. (2022). “Semiotic software through the lens of systemic functional theory,” in *Applicable Linguistics and Social Semiotics: Developing Theory from Practice*, eds. D. Caldwell, J. S. Knox, and J. R. Martin (London/New York: Bloomsbury Academic), 421–435. doi: 10.5040/9781350109322.ch-23
- ELAN (2023). *ELAN-Linguistic Annotator*. Technical report, Max Planck Institute for Psycholinguistics. The Language Archive, Nijmegen, The Netherlands. Computer software Available online at: <https://archive.mpi.nl/ta/elan> (accessed April 1, 2024).
- Fiorella, L. (2021). “Multimedia learning with instructional video,” in *The Cambridge handbook of multimedia learning*, ed. R. E. Mayer (Cambridge, UK: Cambridge University Press), 487–497. doi: 10.1017/9781108894333.050

that could be construed as a potential conflict of interest.

The authors declared that they included an editorial board member of Frontiers at the time of submission. This had no impact on the peer review process and the final decision.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcomm.2024.1356495/full#supplementary-material>

- Fiorella, L., and Mayer, R. E. (2021). "Principles based on social cues in multimedia learning," in *The Cambridge handbook of multimedia learning*, ed. R. E. Mayer (Cambridge, UK: Cambridge University Press), 277–285. doi: 10.1017/9781108894333.029
- Habel, C., and Acartürk, C. (2006). "On reciprocal improvement in multimodal generation: co-reference by text and information graphics," in *Proceedings of the Workshop on Multimodal Output Generation MOG 2007* (Centre for Telematics and Information Technology (CTIT), University of Twente), 69–80.
- Halliday, M. A. K., and Hasan, R. (1976). *Cohesion in English*. London: Longman.
- Halliday, M. A. K., and Matthiessen, C. M. I. M. (2013). *Halliday's Introduction to Functional Grammar*. London and New York: Routledge. doi: 10.4324/9780203431269
- Henderson, M. L., and Schroeder, N. L. (2021). A systematic review of instructor presence in instructional videos: effects on learning and affect. *Comput. Educ. Open* 2:100059. doi: 10.1016/j.caeo.2021.100059
- Holsanova, J., Holmberg, N., and Holmqvist, K. (2008). Reading information graphics: the role of spatial contiguity and dual attentional guidance. *Appl. Cogn. Psychol.* 23, 1215–1226. doi: 10.1002/acp.1525
- Jefferson, G. (2004). "Glossary of transcript symbols with an introduction," in *Conversation Analysis. Studies from the first generation* (Benjamins, Amsterdam), 13–31. doi: 10.1075/pbns.125.02jef
- Just, M. A., and Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychol. Rev.* 87, 329–354. doi: 10.1037/0033-295X.87.4.329
- Kang, S., Tversky, B., and Black, J. B. (2015). Coordinating gesture, word, and diagram: explanations for experts and novices. *Spatial Cogn. Comput.* 15, 1–26. doi: 10.1080/13875868.2014.958837
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychol. Rev.* 95, 163–182. doi: 10.1037/0033-295X.95.2.163
- Kintsch, W., and van Dijk, T. A. (1978). Towards a model of text comprehension. *Psychol. Rev.* 85, 363–394. doi: 10.1037/0033-295X.85.5.363
- Knight, D., and Adolphs, S. (2020). "Multimodal corpora," in *A Practical Handbook of Corpus Linguistics*, eds. M. Paquot, and S. T. Gries (Cham: Springer Nature Switzerland), 353–371. doi: 10.1007/978-3-030-46216-1\_16
- Kosslyn, S. M. (1993). *Elements of Graph Design*. New York, NY: Freeman.
- Larkin, J. H., and Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cogn. Sci.* 11, 65–99. doi: 10.1111/j.1551-6708.1987.tb00863.x
- Levin, J. R., Anglin, G. J., and Carney, R. N. (1987). "On empirically validating function of pictures in prose," in *The psychology of illustration*, eds. D. M. Willows, and H. A. Houghton (New York, NY: Springer), 51–85. doi: 10.1007/978-1-4612-4674-9\_2
- Liu, Y., and O'Halloran, K. L. (2009). Intersemiotic Texture: analyzing cohesive devices between language and images. *Soc. Semiot.* 19, 367–388. doi: 10.1080/10350330903361059
- Lu, J. C.-C. (2023). Using youtube as an effective educational tool to improve engineering mathematics teaching during the COVID-19 pandemic. *Eng. Proc.* 38:24. doi: 10.3390/engproc2023038024
- Martin, J. R., and Unsworth, L. (2023). *Reading Images for Knowledge Building: Analyzing Infographics in School Science*. London and New York: Routledge. doi: 10.4324/9781003164586
- Mayer, R. E. (2009). *Multimedia Learning, chapter Temporal Contiguity Principle*. Cambridge: Cambridge University Press, 153–170. doi: 10.1017/CBO9780511811678.011
- Mayer, R. E. (2014). "Cognitive theory of multimedia learning," in *The Cambridge Handbook of Multimedia Learning, Cambridge Handbooks in Psychology*, ed. R. E. Mayer (Cambridge, MA: Cambridge University Press), 45–71. doi: 10.1017/CBO9781139547369.005
- Mayer, R. E. (2021a). "Cognitive theory of multimedia learning," in *The Cambridge handbook of multimedia learning*, ed. R. E. Mayer (Cambridge, UK: Cambridge University Press), 57–72. doi: 10.1017/9781108894333.008
- Mayer, R. E. (2021b). "The multimedia principle," in *The Cambridge handbook of multimedia learning*, ed. R. E. Mayer (Cambridge, UK: Cambridge University Press), 145–157. doi: 10.1017/9781108894333.015
- Mayer, R. E., Fiorella, L., and Stull, A. (2020). Five ways to increase the effectiveness of instructional video. *Educ. Technol. Res. Dev.* 68, 837–852. doi: 10.1007/s11423-020-09749-6
- Ozcelik, E., Karakus, T., Kursun, E., and Cagiltay, K. (2009). An eyetracking study of how color coding affects multimedia learning. *Comput. Educ.* 53, 445–453. doi: 10.1016/j.compedu.2009.03.002
- Paas, F., and Sweller, J. (2021). "Implications of cognitive load theory for multimedia learning," in *The Cambridge Handbook of Multimedia Learning*, ed. R. E. Mayer (Cambridge, UK: Cambridge University Press), 73–81. doi: 10.1017/9781108894333.009
- Richter, J., Scheiter, K., and Eitel, A. (2016). Signaling text-picture relations in multimedia learning: a comprehensive meta-analysis. *Educ. Res. Rev.* 17, 19–36. doi: 10.1016/j.edurev.2015.12.003
- Rowley-Jolivet, E. (2004). Different visions, different visuals: a social semiotic analysis of field-specific visual composition in scientific conference presentations. *Visual Commun.* 3, 145–175. doi: 10.1177/147035704043038
- Royce, T. D. (1998). *Synergy on the page: exploring intersemiotic complementarity in page-based multimodal text*. Japan Association for Systemic Functional Linguistics (JASFL) Occasional Papers, 25–49.
- Schmidt-Borcherding, F., Bateman, J. A., and Thiele, L. (in preparation). Facing text and graphics in instructional video: The role of instructor presence and coherence signals.
- Schneider, S., Beege, M., Nebel, S., and Rey, G. D. (2018). A meta-analysis of how signaling affects learning with media. *Educ. Res. Rev.* 23, 1–24. doi: 10.1016/j.edurev.2017.11.001
- Schnettler, B., and Knoblauch, H. (2007). *PowerPoint-Präsentationen. Neue Formen der gesellschaftlichen Kommunikations von Wissen*. Konstanz: UVK.
- Schnotz, W. (2014). "An integrated model of text and picture comprehension," in *The Cambridge Handbook of Multimedia Learning*, ed. R. E. Mayer (Cambridge: Cambridge University Press), 72–103. doi: 10.1017/CBO9781139547369.006
- Schnotz, W. (2021). "Integrated model of text and picture comprehension," in *The Cambridge Handbook of Multimedia Learning*, ed. R. E. Mayer (Cambridge, UK: Cambridge University Press), 82–99. doi: 10.1017/9781108894333.010
- Trabelsi, O., Souissi, M. A., Scharenberg, S., Mrayeh, M., and Gharbi, A. (2022). YouTube as a complementary learning tool in times of COVID-19: SELF-reports from sports science students. *Trends Neurosci. Educ.* 29:100186. doi: 10.1016/j.tine.2022.100186
- Tseng, C. (2013). *Cohesion in Film: Tracking Film Elements*. Basingstoke: Palgrave Macmillan. doi: 10.1057/9781137290342
- Tversky, B., Jamalain, A., Giardino, V., Kang, S., and Kessell, A. (2013). "Comparing gestures and diagrams," in *10th International Gesture Workshop, Tilburg* (Tilburg center for Cognition and Communication (TiCC)).
- Tversky, B., Morrison, J. B., and Betrancourt, M. (2002). Animation: can it facilitate? *Int. J. Hum. Comput. Stud.* 57, 247–262. doi: 10.1006/ijhc.2002.1017
- Tversky, B., Zacks, J. M., Lee, P., and Heiser, J. (2000). "Lines, blobs, crosses, and arrows: diagrammatic communication with schematic figures," in *Theory and Application of Diagrams*, eds. M. Anderson, P. Cheng, and V. Haarslev (Berlin: Springer), 221–230. doi: 10.1007/3-540-44590-0\_21
- van der Sluis, I., and Krahmer, E. (2007). Generating multimodal references. *Disc. Proc.* 44, 145–174. doi: 10.1080/01638530701600755
- van Gog, T. (2021). "The signaling (or cueing) principle in multimedia learning," in *The Cambridge Handbook of Multimedia Learning*, ed. R. E. Mayer (Cambridge, UK: Cambridge University Press), 221–230. doi: 10.1017/9781108894333.022
- Wang, J., and Antonenko, P. D. (2017). Instructor presence in instructional video: effects on visual attention, recall, and perceived learning. *Comput. Hum. Behav.* 71, 79–89. doi: 10.1016/j.chb.2017.01.049
- Wiebe, E., Slykhuys, D., and Annetta, L. (2007). Evaluating the effectiveness of scientific visualization in two powerpoint delivery strategies on science learning for preservice science teachers. *Int. J. Sci. Mathem. Educ.* 5, 329–348. doi: 10.1007/s10763-006-9041-z
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). "ELAN: a professional framework for multimodality research," in *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation* 1556–1559.
- Yaacob, Z., and Saad, N. H. M. (2020). Acceptance of youtube as a learning platform during the covid-19 pandemic: the moderating effect of subscription status. *TEM J.* 9, 1732–1739. doi: 10.18421/TEM94-54
- Zhao, S., Djonov, E., and van Leeuwen, T. (2014). Semiotic technology and practice: a multimodal social semiotic approach to PowerPoint. *Text Talk* 34, 349–375. doi: 10.1515/text-2014-0005



## OPEN ACCESS

## EDITED BY

Claudia Lehmann,  
University of Potsdam, Germany

## REVIEWED BY

Manon Lelandais,  
Université Paris Cité, France  
Barbara Dancygier,  
University of British Columbia, Canada

## \*CORRESPONDENCE

Anna Wilson  
✉ anna.wilson@area.ox.ac.uk

RECEIVED 16 December 2023

ACCEPTED 13 February 2024

PUBLISHED 24 April 2024

## CITATION

Wilson A, Pavlova I, Payne E, Burenko I and  
Uhrig P (2024) World futures through RT's  
eyes: multimodal dataset and interdisciplinary  
methodology.  
*Front. Commun.* 9:1356702.  
doi: 10.3389/fcomm.2024.1356702

## COPYRIGHT

© 2024 Wilson, Pavlova, Payne, Burenko and  
Uhrig. This is an open-access article  
distributed under the terms of the [Creative  
Commons Attribution License \(CC BY\)](#). The  
use, distribution or reproduction in other  
forums is permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original publication in  
this journal is cited, in accordance with  
accepted academic practice. No use,  
distribution or reproduction is permitted  
which does not comply with these terms.

# World futures through RT's eyes: multimodal dataset and interdisciplinary methodology

Anna Wilson<sup>1\*</sup>, Irina Pavlova<sup>1</sup>, Elinor Payne<sup>2</sup>, Ilya Burenko<sup>3,4</sup> and  
Peter Uhrig<sup>3,5</sup>

<sup>1</sup>Oxford School of Global and Area Studies, University of Oxford, Oxford, United Kingdom, <sup>2</sup>Faculty of Linguistics, Philology and Phonetics, University of Oxford, Oxford, United Kingdom, <sup>3</sup>Centre for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig, Dresden, Germany, <sup>4</sup>Technische Universität Dresden, Dresden, Germany, <sup>5</sup>Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

There is a need to develop new interdisciplinary approaches suitable for a more complete analysis of multimodal data. Such approaches need to go beyond case studies and leverage technology to allow for statistically valid analysis of the data. Our study addresses this need by engaging with the research question of how humans communicate about the future for persuasive and manipulative purposes, and how they do this multimodally. It introduces a new methodology for computer-assisted multimodal analysis of video data. The study also introduces the resulting dataset, featuring annotations for speech (textual and acoustic modalities) and gesticulation and corporal behaviour (visual modality). To analyse and annotate the data and develop the methodology, the study engages with 23 26-min episodes of the show 'SophieCo Visionaries', broadcast by RT (formerly 'Russia Today').

## KEYWORDS

multimodality, dataset, methodology, speech, gesture, future, conceptual blending, computer vision

## 1 Introduction

This article presents a new methodology for computer-assisted multimodal annotation and analysis of video data and introduces the resulting dataset. The development of this methodology constitutes a stepping stone in our attempt to answer an overarching research question about how humans communicate multimodally about different conceptions of the future for persuasive and manipulative purposes. Manipulation and persuasion constitute propaganda whenever the true intent of the message is not known to the audience (Jowett and O'Donnell, 2006). They are more effective when communicated multimodally (for review, see Wilson et al., 2023).

To analyse and annotate our data and develop our methodology, we engage with 23 26-min episodes of the RT 'interview' show 'SophieCo Visionaries'. We focus on speech (textual and acoustic modalities) and gesticulation and corporal behaviour (visual modality).

We demonstrate our exploratory engagement with the data through a case study of how multimodal cues trigger the construction of meaning, stance, and viewpoint in a hypothetical future depiction by the RT show host (Section 2). The case study does not offer an exhaustive analysis but works to indicate where cues from different modalities are coordinated. It is one of many conducted to shape our approach and enable the development of our methodology and annotated dataset (Section 3). Although our study presents the case study, our

methodology, and our dataset in a linear manner, the processes of conducting case studies, the creation of our annotated dataset, and the development of tools for automated annotation are interdependent and complementary.

Our empirical and data-driven approach is based on the fusion of knowledge and methods from cognitive linguistics, phonetics and phonology, gesture studies, and computer and engineering sciences. We work to find ‘ways of ‘combining’ insights from the variously imported theoretical and methodological backgrounds brought along by previous non-multimodal stages of any contributing discipline’ (Bateman, 2022a, p. 48). We go where the data take us, and do not disregard data that do not fit our hypotheses at the outset of our studies. We consider larger spoken discourse units with their prosodic features and gesticulation as they contribute to viewpoint construction at the semantic-syntactic and pragmatic levels. We rely on technology to speed up and scale up our analysis.

Our multimodal analysis is situated within the framework of conceptual integration/blending theory (Fauconnier and Turner, 2002), which it extends to investigate how multimodal cues—textual, acoustic, and gestural and corporal—trigger the construction of meaning, stance, and viewpoint in RT’s depictions of the future.

Notions of viewpoint and stance are often used interchangeably (Vandelanotte, 2017; Andries et al., 2023). We differentiate between the two, defining viewpoint as a key parameter of a multimodal setup or evoked mental space that represents a point of view of the Speaker or her Interlocutor at this given point in discourse. Viewpoint is ‘marked by just about anything that builds a particular individual’s mental space construal in ways specific to that individual’s cognitive and perceptual access’ (Sweetser, 2012, p. 7). We define stance as epistemic or evaluative constructs in relation to subjects, objects, or states of affairs and as a lower-level phenomenon than viewpoint, while simultaneously influencing configurations of viewpointed mental spaces. We see the viewpoints of the RT host and her guest as voices in Bakhtin’s sense (Bakhtin, 2013). We see their stances as blocks in the building of these voices. We use the term ‘stance construction’ rather than ‘stance-taking’ to reflect its key role in the construction of meaning and viewpoint (cf. Dancygier et al., 2019).

We incorporate in our research insights and methods from prosody and gesture studies, as well as from studies on the interaction of the two (for a literature review, see Loehr, 2014; for recent scholarship, see Pouw et al., 2023).

We use a theoretical approach for prosodic analysis and annotation grounded in the Autosegmental-Metrical approach to intonation (Pierrehumbert, 1980). It sits within a hierarchical theory of prosodic organisation, as expounded by, among others, Nespor and Vogel (1986), Hayes (1989), and Selkirk (2003). We approach our analysis of both prosody and prosody–gesture relations without any prescribed limits to our eventual interpretation, working with all the features together to account for multimodality.

Our interest in the conceptualisations of futures in speech and gesture motivates our interest in temporal gesture (for reviews, see Núñez and Cooperrider, 2013; Cooperrider et al., 2014). We see temporal gesture as belonging to the class of representational gestures, which are defined by Chu et al. as depicting ‘a concrete or abstract concept with the shape or motion of the hands [iconic gestures and metaphoric gestures in McNeill (1992), or point to a referent in the physical or imaginary space (concrete or abstract deictic gestures in McNeill (1992)]’ (Chu et al., 2014, p. 2).

In our analysis of the speech–gesture relation, we draw upon the Information Packaging Hypothesis, which ‘states that gesturing helps the speaker organise information in a way suitable for linguistic expression’ (Kita, 2000, p. 180), with the organisation of information relying on collaboration between the speaker’s analytic and spatio-motoric thinking. We see gestures as communicating information (Hostetter, 2011). We define interactive gestures as referring ‘to the interlocutor rather than to the topic of conversation, and they help maintain the conversation as a social system’ (Bavelas et al., 1992, p. 469).

We treat the questions of what gesture is and what gestural boundaries are as open. We do not have preconceived notions of the direction or form of temporal gestures. We analyse gesture–speech relation in RT shows empirically to offer more complete evidence-based answers to these questions (see Uhrig et al., 2023). Therefore, we adopt the notion of a gestural unit or gestural movement rather than the notion of gesture. We view every gestural movement as potentially carrying more than one function (cf. Kok et al., 2016) and discard preconceived notions of gesture annotation such as phases.

For speech, prosodic, and gestural annotation, we use formal, directly observable categories, following Bateman’s call for the use of external languages of description to avoid the ‘danger of becoming ‘stuck’ within [our] pre-existing conceptualisations’ (Bateman, 2022a, p. 53).

There is a wealth of information in human communication that needs to be annotated to allow for a statistically valid analysis. Beyond the addition of huge amounts of (hu)manpower, the only feasible way to ensure that ‘work at scales larger than individual case studies is to be possible’ (Bateman, 2022a,b, p. 42) is to scale up annotation leveraging technology. Therefore, any annotation scheme must be designed to reflect the needs for analysis as informed by case studies and the affordances and constraints of current computer science and engineering methods.

In leveraging technology to scale up and speed up our research, we work to preserve the fine-grained nature of our analysis wherever possible, thus minimising the associated risk that the detail required will ‘restrict the objects of investigation that multimodality can address’ (Bateman, 2022a, p. 42).

Our computational study is driven by our conceptual thinking. Our conceptual thinking is affected by computational parameters. Both are affected by practical considerations. We determine an optimal interdisciplinary approach and implement it at every stage of our research, which makes our approach novel and our resulting annotated dataset different from other multimodal annotated datasets, in that:

- i the majority of datasets annotated for speech and gesture—with some also annotated for prosody (e.g., Kibrik, 2018)—rely on data collected in experimental (lab) conditions, e.g., SAGA (Lücking et al., 2010), CABB (Eijk et al., 2022), FreMIC (Rühlemann and Ptak, 2023), and Mittelberg (2018). These are not ‘naturally occurring’ data in the sense of Sinclair (1991, 171). In contrast, our annotated dataset is generated using media data, which are regarded by linguists as ecologically valid;
- ii those annotated datasets that have used media data either exercised a fully automatic approach to annotation for gesture generation, e.g., the TED Gesture Dataset (Yoon et al., 2019), or a different manual approach, e.g., Valenzuela et al., 2020, used the NewsScape corpus (Steen et al., 2018) to categorise



temporal expressions co-occurring with gesture, but in contrast to our approach, they did not do a data-driven study, annotate their data in ELAN,<sup>1</sup> or include prosody; and

- iii we have developed computational tools for the automatic annotation of media data and written those annotations into our ELAN files. Our task here was more complex compared to those research teams engaging with lab recordings because of our engagement with media data (e.g., the problem of changes in camera perspective; see Section 3).

## 2 Case study

The case study presents the results of manual analysis of an episode of RT's show 'SophieCo Visionaries', to illustrate the level and nature of detail needed to address our research question and to inform decisions about what kinds of multimodal cues to annotate for in an automatic or semi-automatic annotation scheme. It shows how modalities—textual, acoustic, gestural, and corporal—may work together to construct subtly manipulative messages.

In the video clip A,<sup>2</sup> the host, Sophie Shevardnadze, is in conversation via video conferencing with Tim Kendall, the ex-Facebook Monetisation Director, the ex-President of Pinterest, and the CEO of Moment (United States), about how people have lost control of their smartphones and have become addicted to using social media via them and to scrolling all the time, despite being aware of the associated harmful effects on their health.



Looking at her guest, Sophie produces three multimodal utterances engaging with a hypothetical future depiction. She constructs meaning, stance, and her viewpoint as part of the interaction with her guest to cast doubt upon the validity of her guest's viewpoint. This forms part of a bigger manipulative strategy of discrediting anything that comes from the West and propagating the idea that the West is inferior to Russia in all respects. Making such ideas 'infectious' relies on more than just the multimodal signal produced and received; it relies on various contexts—e.g., situational, linguistic, cultural, and historical—in which the producer and the receiver find themselves. Our case study focuses on determining key cues from three modalities that trigger the construction of meaning, viewpoint, and stance and exploring ways in which the cues are coordinated in the video clip to prompt the audience to share Sophie's viewpoint.

The guest is not visible in the clip under examination, but he appears on screen either by himself or simultaneously with Sophie elsewhere in the show. As is normal for TV broadcasting, Sophie's audience is both her interlocutor (guest) and the TV audience (the implied viewer). The audience is prompted to construct a scene of blended joint attention, in which Sophie and them are attending jointly to the topic about smartphones and social media (Turner, 2014, p. 97–105).

As part of this scene (see Table 1 below for visual representation), Sophie says:

*Wait so you are the CEO of Moment now—an app which according to the description of the website helps people build healthier relationships with their phones. So if I delete all social networks from my phone, how will my relationship with it become healthier exactly? I mean because, you know, I can really just check Twitter on desktop.*

To analyse this example, we utilise conceptual integration/blending theory (Fauconnier and Turner, 2002), also making use of several tools and insights from mental spaces theory (Fauconnier, 1994). These are two related cognitive theories of meaning construction that are often drawn upon for the analysis of persuasive and manipulative discourse (see Pleshakova, 2018 for review). We also rely on the 'mental spaces' analysis of causal and conditional conjunctions by Dancygier and Sweetser (2000, 2005). Their studies demonstrate that conditionals like *if* and *because* can set up various mental spaces while fulfilling various communicative functions. *If* can introduce patterns of reasoning at different levels (e.g., predictive, epistemic, or metalinguistic); it can build epistemically distanced or non-distanced or neutral spaces; and those spaces can then be referred to deictically. Dancygier and Sweetser (2005, p. 58) differentiate between non-conditional, positive-stance future predictions, which 'are about an expected future (unrealized) development of reality', and conditional, negative-stance future predictions, which are about future 'not yet realised and not certain to be realised'. They argue that:

*[...] if [...] expresses the speaker's lack of full positive stance with respect to the content. The non-positive stance of if need not commit the speaker to a negative or sceptical stance, but does indicate that she thereby distances herself from full commitment to the contents of the if-clause. Other aspects of a conditional construction may go further, and explicitly mark the speaker's leaning towards non-belief in the reality of the described situation (Dancygier and Sweetser, 2000, p. 125).*

Space-building functions of *because*-clauses are different, as 'causal conjunctions are semantically more appropriate to elaboration of spaces' (Dancygier and Sweetser, 2005, p. 172, 181). The authors showcase the complexity of the mappings between information structure, clause order, and expressions of conditional and causal relationships (Dancygier and Sweetser, 2005, Ch. 7). The human mind is embodied, and we extend the framework of conceptual integration/blending to investigate not only how language and gesture work together in meaning and viewpoint construction (e.g., Parrill and Sweetser, 2004; Narayan, 2012; Parrill, 2012; Tobin, 2017; Turner et al., 2019; Valenzuela et al., 2020), but also how cues of speech (including prosody), gesticulation, and corporal behaviour trigger the construction of meaning, stance, and viewpoint in manipulative media communication.

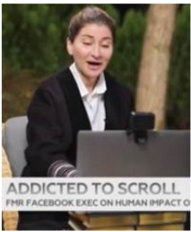
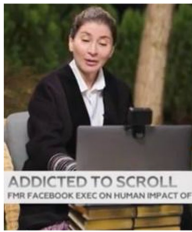

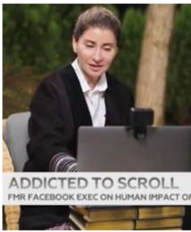
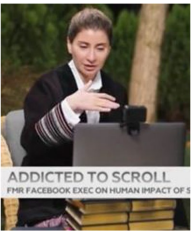
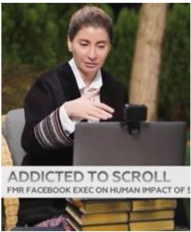
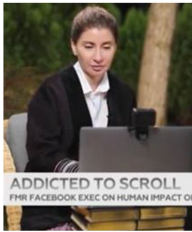
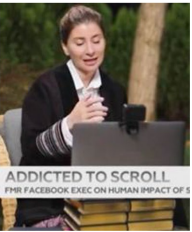
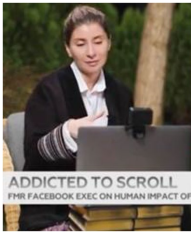


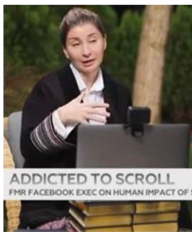








### 2.1 Multimodal triggers at work: mental space 'M'

The RT host Sophie engages with what she herself presents as the viewpoint of her guest: *Wait so you are the CEO of Moment now—an app which according to the description of the website helps people build healthier relationships with their phones*. She cites the description on the company's website and states that her guest is the CEO of the

<sup>1</sup> <https://archive.mpi.nl/tla/elan>







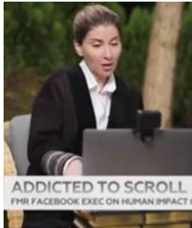
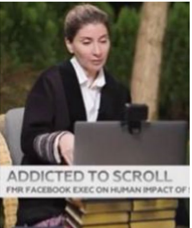
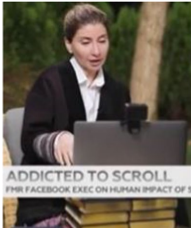
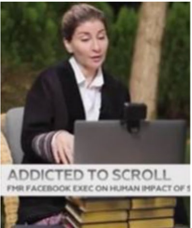
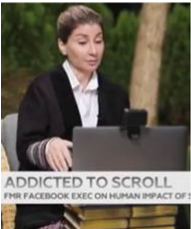
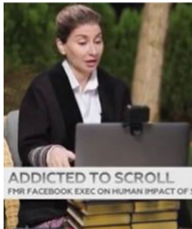
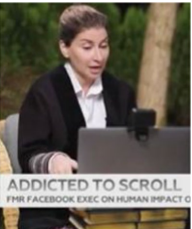
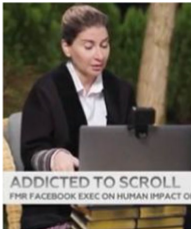
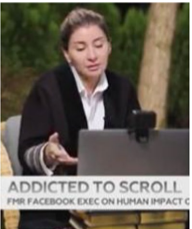
<sup>2</sup> See <http://go.redhenlab.org/pgu/0132/> or scan the QR code.

TABLE 1 Multimodal utterances presented in stills.

A1	A2	A3	A4	A5
Wait so			you are the CEO of Moment now, an app	which according
				
Eyebrow U <sup>1</sup>	Head tilt L eyebrow U	Head tilt L eyebrow D	Head tilt F D	RH U fingers L, thumb U
A6	A7	A8	A9	A10
to the	description of the website	helps	people	
				
RH U, fingers F, throwing and shaking movements F	RH D	RH U, head tilt R	RH and RH fingers D and F beat, thumb L	RH and RH fingers U and body-directed, thumb U
A11	A12	A13	A14	A15
build	healthier	Relationships		
				
RH and RH fingers D, RH circular movement	RH and RH fingers F, RH circular movement R, head tilt R	RH L and then R and U, fingers U, head D	RH D and L	RH D, head tilt U
A16	A17	A18	A19	A20
with	their phones		Uhhhh	So
				

(Continued)

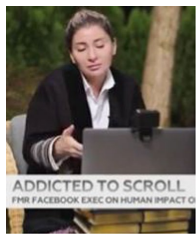
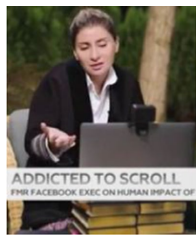
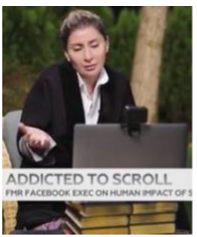
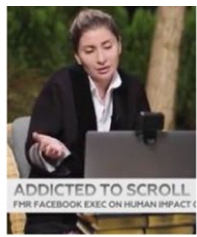
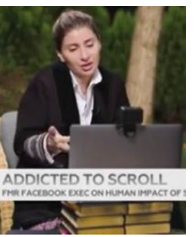
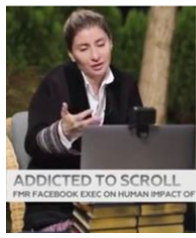
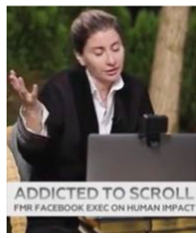
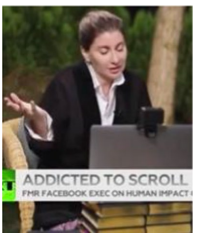
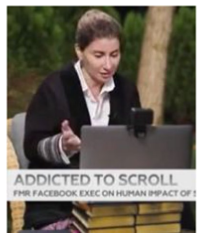
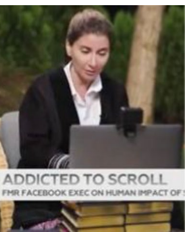
TABLE 1 (Continued)

corporal movement F (right side)	RH holding the phone U and then D	The phone is back on the desk. Head nod. Closed eyes. [Tim backchannels 'yeah' in confirmation]	Gaze L (not focused on the screen)	RH U, handshape 'phone'
A21	A22	A23	A24	A25
If I	delete	all	social networks	from my phone
				
RH R, slicing gesture	RH R and D, slicing gesture	Hands not visible, eyebrows U	2 prosodic words Hands not visible; eyebrows U	Hands not visible;
A26	A27	A28	A29	A30
	How will	my relationship		with it
				
Hands not visible; eyebrows U	RH U	RH U	RH shaking (L-R)	Body lifts up a bit, RH R, shaking (L-R), eyebrows U, head U
A31	A32	A33	A34	A35
become	healthier	exactly		I mean
				
RH R and D beats/shaking, eyebrows U	RH R and D, beats/shaking, eyebrows U	Hold, leaning F, eyebrows U	Hold, eye blinking	Shoulder shrug and head tilt R, RH rotates at wrist, RH OPU U and F
A36	A37	A38	A39	A40
[mean] ['cause]	'cause	[you know I] can	really	just

(Continued)



TABLE 1 (Continued)

				
RH U-L, fingers U and L, shoulder shrug and head tilt R	RH R, shoulder shrug and head tilt R	RH R, shoulder shrug and head tilt R	RH L, shoulder shrug and head tilt R	RH L (C), shoulder shrug and head tilt R
A41	A42	A43	A44	A45
check	Twitter	on desktop		
				
RH U and R, shoulder shrug and head tilt R	RH R and U, shoulder shrug and head tilt R	RH R and D, shoulder shrug and head tilt R	RH L and slightly F, shoulders down and head straightens	RH L and D, shoulders down and head tilt L

<sup>1</sup>Section 3.3 for glossaries of abbreviations.

company, thereby implying that the website ultimately conveys her guest's viewpoint. Sophie's verbal statement and her prosodic, gestural, and corporal behaviour (A1–A18) work to set up the viewpointed mental space (M), which incorporates:

- i Base: Sophie and her guest in interaction, the context of the RT show;
- ii Content: Sophie's guest is the CEO of Moment—an app described on the company's website as helping people build healthier relations with their phones;
- iii Focus: the description on the website as the guest's viewpoint;
- iv Sophie's Epistemic Stance of certainty towards the Content of the website's description and her guest's occupation and the link between them;
- v Sophie's potentially negative Evaluative Stance towards the guest's contribution made immediately before .<sup>3</sup>

Sophie's prosodic behaviour on *wait so* and facial gestures—smiling and eyebrows moving up (a 'peak')<sup>4</sup> (A1–A3)—signal her surprise at the content of her guest's contribution immediately before. The accompanying prosody and facial gestures help to manage interaction at this turn-taking point. Both make *wait so* more prominent.

She produces each word—*wait* and *so*—as individual phrases. There is strong marking of the final boundary of each of those phrases,

with strong glottalisation at the end of both (Figure 1). These two short phrases, which we have interpreted as intermediate phrases (ip), have their own nuclear pitch accents, and together they form a somewhat rhythmic pattern, perceptually. Sophie also starts to say something else, beginning with [w], which could be interpreted as the start of a *wh*-question before *you are*. She then reconfigures what she wants to say. The effect is a strong signalling of 'hold on a second...', and therefore manifests questioning and sceptical stance.

Sophie's smile signals that she has spotted incongruity in her guest's contribution and that she may doubt credibility behind his viewpoint.

Sophie's right-hand gestures co-occurring with *you are the CEO of Moment now—an app which according to the description of the website helps people build healthier relationships with their phones* are performed in the central gestural zone.

On *you are the CEO of Moment now—an app which according to the description of the website*, Sophie engages the vertical, lateral, and sagittal axes (A4–A7) to conceptually map the description on the website to her guest. The fingers of her right hand go forward in a quick throwing move to represent the mapping. In addition to the representational function, this movement carries an interactive function in helping to maintain the dialogue between Sophie and her guest.

On *helps people build healthier relationships*, Sophie performs a complex sequence of right-hand gestural movements of various amplitudes and performed at a changing pace. This complex gestural configuration engages vertical, lateral, and sagittal axes to depict the non-straightforward process of the building of the healthier relationship (A8–A15). On *with their phones*, Sophie's right hand goes down to pick up her phone and show it to her guest before putting it back on the desk (A16–A18). Following a quick smile at the beginning

3 See <http://go.redhenlab.org/pgu/0137> or scan the QR code.

4 See Section 3.3.2.3 for explanation.



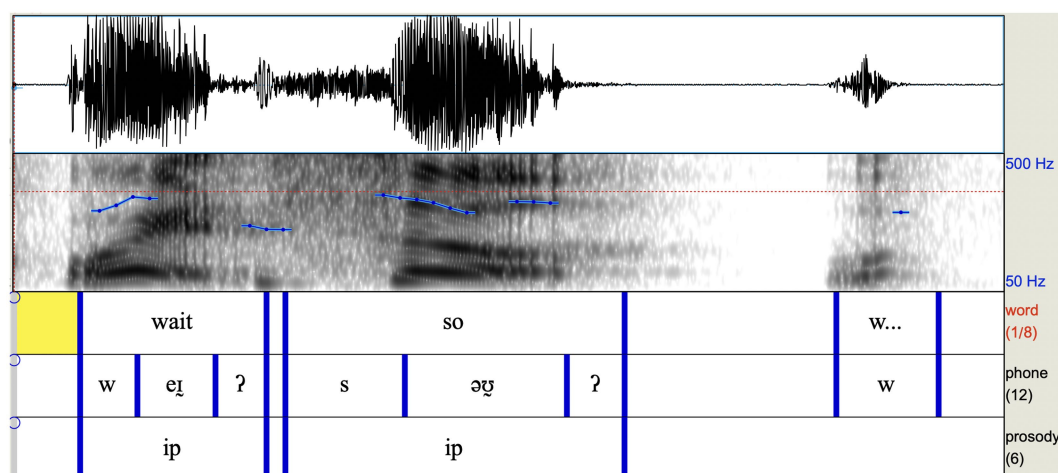


FIGURE 1  
Waveform and spectrogram for *wait... so...* with segmental and prosodic annotation (intonational phrases, pitch accents).

of the utterance, Sophie's facial expression remains neutral throughout, her gaze is focused on the screen. Sophie closes her eyes at the end of her first utterance (A18). There is a simultaneous head nod against the background of her guest's backchannelling *yeah* serving as further confirmation of the accuracy of Sophie's representation of her guest's viewpoint.

## 2.2 Multimodal triggers at work to enable conceptual and viewpoint blending

The setup of the mental space M1 relies on the mental space M as the input. The mapping between the viewpointed spaces M and M1 enables the construction of the conceptual and viewpoint blending network and the emergence of the new mental space M2, representing the viewpoint blend.

The network construction is triggered by Sophie’s next multimodal utterance—*So if I delete all social networks from my phone, how will my relationship with it become healthier exactly?*

Space M1 incorporates:

- i Base: Sophie and her guest in interaction, and the context of the RT show.
- ii Content: the hypothetical future scenario.
- iii Focus: the hypothetical future scenario—*if*-clause—presented by Sophie and the *how*-question about Sophie's future relationship with her phone becoming healthier.
- iv Viewpoint and Epistemic Stance of uncertainty as pertinent to hypothetical future scenarios (the Speaker distances herself from full commitment to the content of the *if*-clause).

Before Sophie utters the *if*- and *how*-clauses, she says *uhhh* and her gaze goes left signalling her collecting her thoughts (Brône et al., 2017). That ‘leftwards—not in focus’ gaze behaviour co-occurring with *uhhh* triggers the process of setting up a new input mental space, M1. Sophie’s multimodal *if*- and *how*-clauses work to configure this new mental space, representing her own viewpoint on the content. M1 is mapped onto space M, which represents the guest’s viewpoint on

the content. The mapping starts the blending process for the two viewpoints—Sophie's and the guest's—thereby supporting the interpretation of Sophie's *if*- and *how*-clauses not as independent units but as part of unfolding discourse—a continuum. The blending process generates the viewpointed blend space of M2, in which M1 is interpreted in relation to M, and the viewpoint of M1 is conceptually presented as more authoritative. M1 as blended with M in M2 is also interpreted in relation to a number of other viewpointed mental spaces set up by the preceding discourse. For example, earlier in the discourse, the guest talks about people being digitally addicted. He talks about people going on their phones to check the weather and realising 45 min later that they have been scrolling through their Facebook news feed or Twitter.

The construction of the blend M2 is already triggered by Sophie's uttering *so in So if I delete all social networks from my phone*. This works to map the content and viewpoint of space M1 to the content and viewpoint of space M. The outer-space mappings are selectively projected into M2 to become the blend's inner-space conceptual relations. The question *how will my relationship with it become healthier exactly?* relies on the presupposition of the predicted result that deleting will lead to a healthier relationship with the host's phone. The latter in turn relies on the input mental space M. This presupposition enables the construction in M1 of the causal relation in the content of the utterance between the deletion and the relationship becoming healthier. Simultaneously, it enables the construction in M1 of the causal relation between the hypothetical event of the deletion and the *how*-question as part of the speech interaction scenario.

Blend space M2 incorporates:

- i Base: Sophie and her guest in interaction; the context of the RT show.
- ii Content: Sophie's guest is the CEO of Moment—an app described on the company's website as helping people build healthier relations with their phones.
- iii Content and Focus: the hypothetical future scenario—if-clause (deletion, phone)—presented by Sophie and the *how*-question about the future (relationship becoming healthier, exactly) that she asks.

- iv Epistemic Stance of uncertainty (the Speaker distances themselves from the content).
- v Evaluative Stance of scepticism.
- vi Alternative 'hypothetical' Content and Focus: the predicted result of the relationship between the Speaker and the phone *not* becoming healthier following the deletion of social networks from Sophie's phone.
- vii Sophie's Viewpoint that deleting social networks from one's phone will not make their relationship with their phone any healthier.

The *if*-utterance comprising *if*- and *how*-clauses is loaded, and by the time Sophie has uttered the *if*-utterance, it is clear that she does not believe that the deletion of the social networks from her phone will make her relationship with her phone any healthier. The whole *if*-utterance is therefore ultimately interpreted through the lenses of the evaluative and epistemic stances in M2, to where the predicted result of the relationship between the Speaker and the phone *not* becoming healthier is projected. This predicted result is dependent on the content of the social networks being deleted and constitutes the alternative to the presupposition that the relationship will become healthier.

Next, Sophie produces the utterance incorporating the *because*-clause: *I mean because, you know, I can really just check Twitter on desktop*.

This utterance triggers the setup of mental space M3 to offer Sophie's reasoning in support of her stance and viewpoint already constructed in M2. Although her argument shifts the focus from her relationship with her phone to the use of social networks more generally, the way she presents this *because*-utterance multimodally creates the impression that she reasons about her relationship with her phone.

Space M3 incorporates:

- i Base: Sophie and her guest in interaction; the context of the RT show.
- ii The Contents of Sophie's interaction with the guest—'I mean, 'you know'—as well as of Sophie's ability to check Twitter on her desktop.
- iii Focus on the reasoning—*because*-clause and making it 'shared' reasoning.
- iv Sophie's Viewpoint—deleting social networks from one's phone does not prevent them from checking social networks on one's desktop.
- v Stance of epistemic certainty.

Mental space M3 is mapped into M and M1 and works to further reconfigure and elaborate the blend space M2. The reconfigured M2–M2(1) presents the *how*-question as expository and as argumentative strategy (see, e.g., Pascual, 2014; Xiang and Pascual, 2016). It features Sophie's epistemic stance of certainty in support of her reasoning (*because*-clause in focus). Her evaluative stance is more openly sceptical. Sophie's reasoning works to further construct her viewpoint that 'deleting social networks from one's phone will not make their relationship with their phone any healthier'. Her viewpoint is constructed as more authoritative and believable, despite the lack of logic in her argument (checking Twitter on her desktop might still make her relationship with her *phone* healthier).

The reconfigured M2(1) blend space incorporates:

- i Base: Sophie and her guest in interaction; the context of the RT show;
- ii The Contents of (a) the situation in which Sophie's guest is the CEO of Moment—an app described on the company's website as helping people build healthier relations with their phones; (b) the hypothetical future scenario—*if*-clause (deletion, phone)—presented by Sophie and the *how*-question about the future (relationship becoming healthier, exactly) that she asks; (c) Sophie's interaction with the guest—*I mean, you know*—as well as of Sophie's ability to check Twitter on her desktop, offered in the form of the *because*-clause.
- iii Focus on the hypothetical future scenario—*if*-clause—presented by Sophie and the *how*-question about the future that she asks.
- iv Focus on the reasoning—*because*-clause and making it 'shared' reasoning.
- v Epistemic Stance of uncertainty (the Speaker distances themselves from the content).
- vi Epistemic Stance of certainty in the 'reasoning' part—the *because*-clause.
- vii The Evaluative Stance of scepticism.
- viii Alternative 'hypothetical' Content and Focus: the predicted result of the relationship between the Speaker and the phone *not* becoming healthier following the deletion of social networks from Sophie's phone.
- ix Sophie's reasoning works to enhance her Viewpoint that 'deleting social networks from one's phone will not make people's relationship with their phone any healthier'. It is constructed to be more authoritative and believable.

## 2.3 Multimodal triggers at work: zooming in

On *so*, Sophie makes a gesture with her right hand to activate the concept of the phone in M1. Her eyes are closed, which may signal the start of the next construction of meaning and viewpoint (A20). The *if*-clause which follows launches the configuration of M1 as a hypothetical future scenario in which Sophie deletes all social network applications from her phone and checks Twitter on her desktop.

The future deletion is conceptualised in gesture through the 'slicing' right-hand rightward and downward movement. The gestural conceptualisation is already there on *if I* (A21) before Sophie utters the verb *delete*. It continues on *delete* (A22).

The *if*-clause comprises seven prosodic words—*So|if I|delete|all|social|networks|from my phone* (Figures 2, 3). There are five pitch accents on *if I*, *delete*, *all*, *networks*, *from my phone* as well as two phrase accents on *delete* and *phone*. The *if*-clause's boundaries co-occur with the boundaries of the intonational phrase (IP), which in turn incorporates two intermediate prosodic phrases—*so if I delete* and *all social networks from my phone* separated by a pause. The nuclear pitch accents within the respective intermediate phrases (ip) fall on *delete* and *from my phone*. The latter two are the only

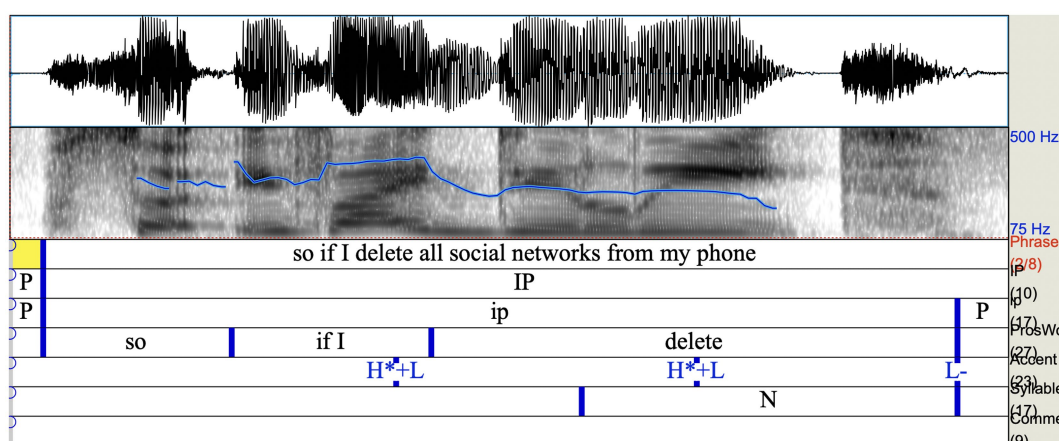


FIGURE 2  
Waveform and spectrogram for *so if I delete* with prosodic annotation (intonational phrases, pitch accents, and pauses).

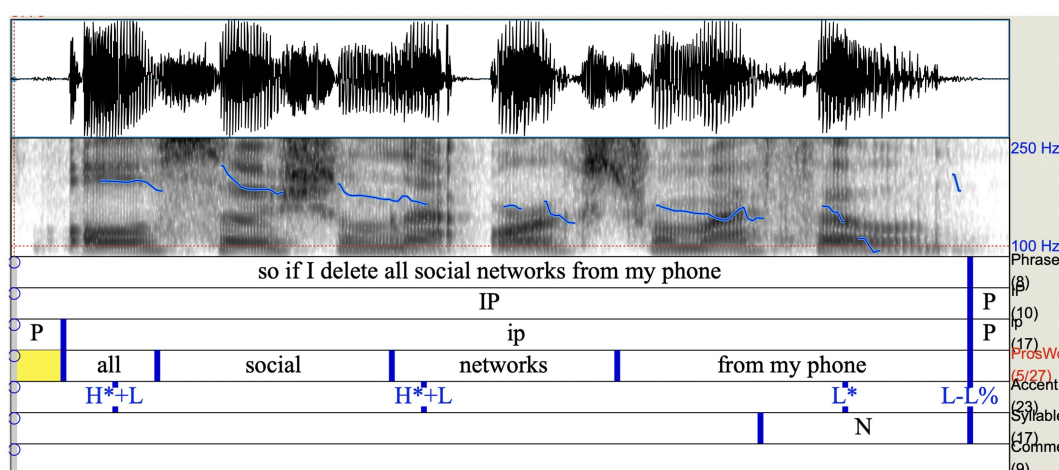


FIGURE 3  
Waveform and spectrogram for *all social networks from my phone* with prosodic annotation (intonational phrases, pitch accents, and pauses).

concepts which are also depicted in hand gesture. Sophie further foregrounds *delete* prosodically via a clear and audible release of the final [t].

At the same time, she puts some prominence on *phone* in speech. It comes at the end of the first IP—and is accompanied by low fall with creaky voice signalling a complete conceptual unit in itself, though it is not the final thing Sophie has to say.

While speech and gestural representations for ‘delete’ co-occur, the hand depiction for ‘phone’ and the speech unit *phone* do not. The gestural *phone* co-occurs with *so* at the very beginning of the *if*-clause and the IP (A20). The speech representation for ‘phone’ is at the very end of the multimodal *if*-clause and the IP (A25). Thus, speech- and hand-gesture triggers for activation of the same concept *phone* are located at the boundaries of the multimodal *if*-clause. Between ‘phone’ in gesture and *phone* in speech, they bookend the whole IP. This multimodal configuration ensures that the concept of ‘phone’ is in focus throughout the clause.

Both prosody and gesture work in a complementary manner to support the configuration of the blend space M2, which, among other things, seems to include an internal hierarchical structure signalling which concepts are more in focus than others. By using the gestural ‘phone’ and *phone* in speech at the edges of the IP, Sophie is constructing the background story as being about the phone. She then has the freedom and flexibility to highlight something else within the sub-structure of the IP. This she does by using the longer-lasting gestural ‘delete’ over the intermediate phrase and having strong emphasis on the word *delete*. Such a distribution of speech–gesture representations for ‘phone’ and ‘delete’ signals the multimodal conceptualisations for the ‘phone’ as fulfilling the ground function and for ‘delete’ as fulfilling the figure against the ground function (On the gestalt psychological principle of figure and ground and the use of the relation in cognitive linguistics, see, e.g., Ungerer and Schmid, 2013, p. 163–191) We also note a parallel in prosodic analysis with the relationship between prosodic ‘domain’—i.e. the prosodic constituent

within which a prosodic feature applies—and prosodic ‘prominence’—i.e. the focal element of the domain in question. Prosodic cues may signal both the boundaries of a domain/constituent (‘edges’) and the focal points within domains (‘heads’).

Due to restricted visibility, we cannot see whether there is any hand gesture performed on the rest of the *if*-clause (A23–26). However, we can see the ‘eyebrow up’ movement co-occurring with *all social networks from my phone*. This relatively long gestural movement—‘plateau’<sup>5</sup>—introduces the stance of wondering, further supporting (i) the conceptualisation in which the RT host distances herself from the depicted future scenario; (ii) the construction of epistemic stance of uncertainty and evaluative stance of scepticism; and (iii) the construction of Sophie’s viewpoint where her deletion of social networks from her phone does not lead to having a healthier relationship with it.

As the *how*-clause is uttered by Sophie, her right hand is formed as a brush with the fingers pointing down (A27–A34). It goes upwards, reaching its highest position on *relationship with it become* (just above the waist level, see A29–A31). On *my relationship*, the RH makes shaking movements. The hand then moves rightwards very slightly on *with it* while still making shaking movements. There is also a slight corporal movement and head to the right (A30). On *healthier*, the right hand goes slightly downwards (A32) and holds the position on *exactly* (A33 and A34). All gestural movements have very small amplitudes. The hand is relaxed, and the fingers are spread. The hand moves upwards and rightwards slightly to mark the future on *will my relationship with it* but remains in the central gestural zone.

There is a contrast between the hand gesture representations of the healthier relationship with one’s phone in the first multimodal utterance (A10–A17) and in the second utterance (A27–A34). Not only do gestural configurations differ in the amplitudes and levels of confidence, but their positioning is also much higher in the first utterance. The direction, including the orientation of fingers, in the first utterance is predominantly upwards–rightwards, whereas in the second utterance it is predominantly downwards–rightwards. Even when the right hand in the second utterance goes upwards, it does not go as high as in the first utterance, and the wrist leads on this ascending in the second utterance with fingers pointing down. This contrast between two gestural configurations co-occurring with the two speech utterances translates into a difference in epistemic and evaluative stance between the two as presented multimodally.

The epistemic and evaluative stances of the *because*-utterance are positive, and the gestural configuration works to communicate that (A36–A45). The overall characteristics of the gestural movements of the *because*-utterance resemble those of the first utterance in that they are of a bigger amplitude, more confident, and the palm orientation is up. The overall direction of the gestural sequence forming part of the *because*-utterance is upwards and rightwards, the same as we observe for the first utterance (cf. A5–A17).

On the prosodic side of the *how*-clause, there is phrase-initial strengthening on the [h] of *how*. This could signal the uncertainty embedded in the question. The nuclear accents in the *how*-clause fall on *my relationship*, *with it*, and *exactly*. In the hand gesture co-occurring with these speech units, we see marking of prominence,

too—the right hand is in an elevated position and shaking on *my relationship*. It is at its highest position and shaking and goes slightly rightwards on *with it*. It is at its lowest position and holding on *exactly*. The three nuclear pitch accents constitute the cores of three intermediate phrases, which in turn form one intonational phrase (IP). The nuclear accents on the speech units *my relationship* and *exactly* create boundaries of the multimodal ground, which in gesture manifests itself through shaking throughout, consistent small amplitude of hand gesture, slight head tilt right and forward, shoulders slightly lifted. At the same time, this ground constitutes the figure of the IP of the *how*-clause as a whole. Sophie creates this multimodal ground/figure to signal the content in focus, which should be evaluated through the prism of the epistemic stance of uncertainty and of the evaluative stance of scepticism in M2. She further foregrounds *with it* as a figure by making a significant pause, thereby placing it in its own intermediate phrase (ip). Furthermore, she uses several phonetic devices to audibly strengthen the ip onset, namely the re-articulation and lengthening [w], as well as articulatory strengthening in the form of ‘stopping’ (the release of which is evident in the spectrogram). Sophie is effectively placing prosodic ‘scare quotes’ around *with it*, thereby distancing herself from the phone and placing it in some kind of isolated relief. She conveys a lack of trust, signalling that she does not really believe one can have a relationship with a phone, or at least not a natural, healthy one (Figure 4).

On *with it become healthier exactly*—the last two intermediate phrases of the *how*-clause—we observe another ‘plateau’ eyebrow gesture and a corporal movement forward (A30–A34). Perceived together, they simultaneously fulfil the functions of ground and of figure in their own right. These gestural and corporal movements, on the one hand, create the ground for figures *with it* and *exactly*, and on the other hand put the unit *with it become healthier exactly* in focus as a figure against the ground of the *how*-clause, working to further configure the hierarchical structure of the M2 blend space.

The eyebrow gestural movement conveys the stance of wondering, which is primarily applied to the content of *with it, become healthier exactly*. The simultaneous corporal movement forward adds to the prominence of this content, and signals Sophie’s intention to really convey this to her guest.

Sophie’s communicative goal is further evident in her phrasing of what follows, separating *I mean*, and ‘*cause you know*’ into intermediate phrases. By isolating first herself and then her interlocutor, she cultivates a knowing and equal ‘pact’ with her interlocutor (i.e., communicating ‘we both know this...’) (Figure 5).

Simultaneously, we observe the dominance of a conduit gestural movement—right hand palm-up going forward—in the *because*-utterance (A35–A41). The movement serves the interactional and representational function of offering content to the interlocutor. It has a special configuration going rightwards in addition to going forward. This rightward movement conveys the temporal function of future depiction. This hand gesture is accompanied by the shoulder shrug and head tilt to the right, which also contribute to the construction of the epistemic and evaluative stance of ‘I am confident that I am right, and I am wondering what objections you can possibly have’. The small-scale move rightwards by the right hand on *can* is in line with its epistemic stance of less certainty (A38). The latter transforms into certainty immediately after, when Sophie’s right hand goes briefly to the centre on *just* (A40) and then goes much further rightwards and upwards on *check Twitter on desktop*. We observe a nuclear accent on

<sup>5</sup> See Section 3.3.2.3 for further explanation.



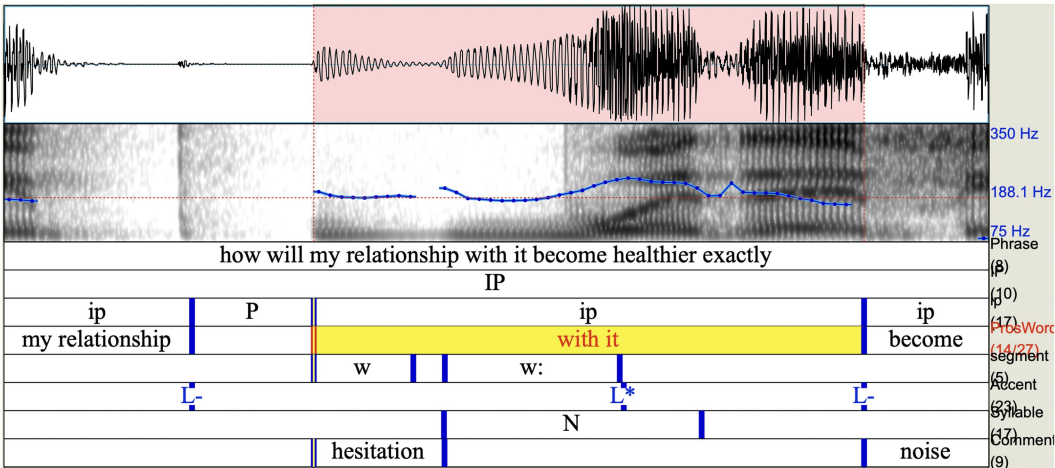


FIGURE 4 Waveform and spectrogram showing the prosodic boundaries around *with it* (in yellow) and re-articulation and strengthening of phrase onset [w] (in pink).

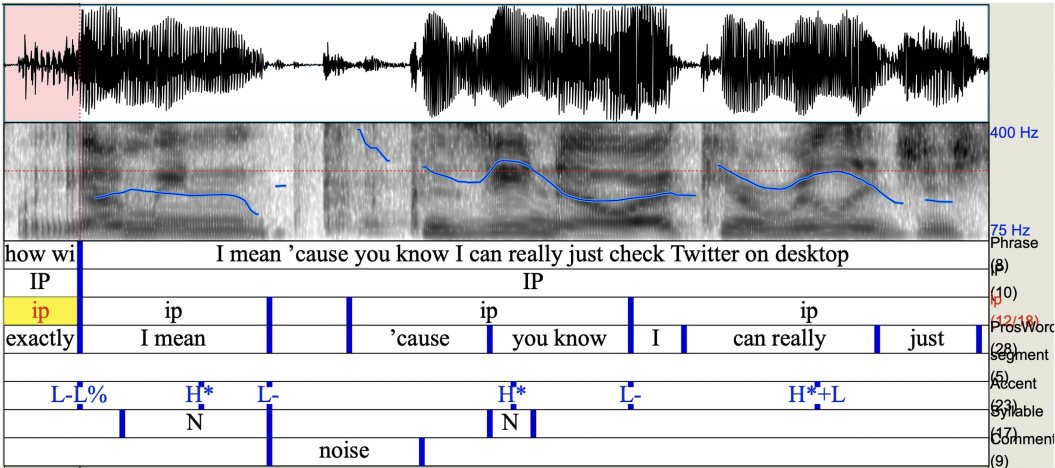


FIGURE 5 Waveform and spectrogram illustrating the intermediate phrasing of *I mean* and *'cause you know*.

*desktop*. At this point, Sophie's right hand is already returning to the centre from its rightmost and highest position on *Twitter* (A42).

There is a quick, repeated eye-blinking and a quick head nod on *desktop* too (A43–A45). The more confident gestural movement—with a bigger amplitude—rightwards and upwards towards the end of the *because*-utterance signals Sophie's belief in this possible future scenario as juxtaposed to the previous future scenario depicted multimodally in the *if*-utterance. Nuclear accents falling on the *desktop* in the *because*-utterance and on the *phone* in the *if*-utterance seem to also serve a special function here linking and juxtaposing *phone* and *desktop* at the same time.

Eye closing plays its own role throughout the three multimodal utterances under consideration (A1–A18; A20–A26; A34–A43). It further marks the boundaries of bigger units (usually IPs), which trigger the construction of meaning in the underlying conceptual blending network.

Core to the network is the emergence of the blend space  $M2 > M2(1)$ , which features an epistemic and evaluative stance of scepticism and disbelief towards the matter of building a healthier relationship with one's own phone.

Using all three modalities in concert, Sophie conveys scepticism about both the phone itself, the possibility of having a relationship with the phone, and also the ability to delete social media from it. She conveys her scepticism multimodally to the interlocutor and the TV audience. We observe a hierarchical interplay of cues—features and phrase boundaries—across the textual, visual, and acoustic modalities; conceptually, the cues play distinct but complementary roles.

One aspect of prosodic structure may be cued by many different acoustic-prosodic cues (and combinations thereof), and at the same time, any given acoustic-prosodic feature can cue different aspects of prosodic structure. This means that there is a non-simplistic association between structure and phonetic implementation in both

directions. We observe the same complex relationship between gestural movements and the underlying gestural structure. We hypothesise that this complex relationship might also be found between gestural and prosodic modalities.

Not only does Sophie use the interplay of multimodal cues to structure communication through a configuration of mental spaces underlying it, she also segments communication to package information, foreground and background pieces of information, and construct her stance towards them. This process ultimately enables the construction of her own viewpoint, which is communicated as more authoritative and believable than her guest's, manipulating the viewer to accept it despite some inherent failures of logic within it.

Our blending analysis demonstrates the importance of engaging with units of various lengths and forms belonging to all three modalities—textual, acoustic, and visual—and their interaction for the study of human communication, including manipulation.

To generalise and hence further develop our approach to manual analysis of our video data, we need to be able to analyse more than one example. To achieve that, we need to leverage technology, and to do that in a well-informed and optimal way, we need to do more case studies, with each contributing to our understanding of the conceptual, computational, and practical aspects of ongoing research. This summarises the iterative process we have gone through to make decisions on annotation and the creation of tools, to design our new ELAN annotation scheme with its meta-language, and to construct the expertly annotated dataset described in Section 3. Several case studies, like the one offered in this section, played an integral part in the development of our new methodology for multimodal analysis presented next.

### 3 Dataset and the development of methodology

In this section, we describe all the levels of annotation in our dataset, thereby presenting our annotation scheme as a whole. We discuss our motivations—conceptual, computational, and practical—for choosing specific annotation levels and values throughout. We describe the way in which we have interwoven manual and computational approaches to annotation. We present our methodological approach to exploratory analysis and simultaneous annotation of ecologically valid multimodal data, which allow to do both on a larger scale and relatively faster. At the initial stage of our study, we explored RT talk shows in English, namely: SophieCo Visionaries, hosted by a woman, Sophie Shevarnadze; and News with Rick Sanchez, hosted by a man, Rick Sanchez. We also examined the four-episode Russian-language documentary on post-Covid futures, *Мир после [The World After]*, hosted by Tina Kandelaki. Having done some preliminary 'speech-gesture' analysis and annotation in ELAN, we opted for first studying SophieCo Visionaries in more depth. This show was of immediate interest to us because it was broadcast by RT in English and its thematic focus was exclusively on world futures. The show constituted data most suitable for answering our research questions, which are centred around the construction of future depictions multimodally for persuasive and manipulative purposes and targeting international audiences. At the outset of our study, we identified 'will' as one of the most frequent speech markers for future depictions. We created a corpus of 20-s video clips centred around 'will' using searches in QQPweb and subsequently analysed 84

clips using the Rapid Annotator<sup>6</sup> to get a preliminary understanding of gestural behaviour of the Speaker co-occurring with future depictions in speech. We subsequently focused on 47 clips in which Sophie was the Speaker and moved to annotating in ELAN to allow for capturing rich multimodal data for more features and in a more precise manner. As we were designing our annotation scheme in ELAN, we had to make several decisions to allow for the annotation to be focused on the 'future' aspect of multimodal depictions, be optimal in terms of labour and time required, and be well balanced in terms of conceptual and computational motivations.

We identified discourse units of various lengths centred around speech markers that trigger the construction of viewpoint future depictions. Those included syntactic clauses, sentences, or even sequence of sentences. We then annotated for gestural sequences co-occurring with those discourse units. We prioritised annotating for sequences of gestural movements that were impressionistically perceived by coders as conceptualising time as a line and motion along the line (e.g., Núñez and Cooperrider, 2013 and Cooperrider et al., 2014). We regarded as open the question of direction and axis for future vs. past vs. present gesture, or, in other words, we refrained from assuming that in English, the future is conceptualised via forward and rightward hand gestures only, and the past is conceptualised via backward and leftward gestures only (cf. Valenzuela et al., 2020). To maintain our focus on the future aspect and to keep annotation manageable and machine-learning friendly, at stage 1, we did not include annotation for iconic gestures such as the 'phone' hand movement discussed in Section 2. This is because it lacks an obvious temporal function. However, we included iconic gestures such as 'delete' as it clearly carries the temporal function of the future in addition to the iconic function of deleting.

The length of intervals chosen for annotation at speech and gestural tiers was determined by our focus on the temporal aspect of meaning and viewpoint construction, as well as practical considerations. Although we had to limit the intervals we could annotate for manually at the first stage, now that we have developed computational approaches for automatic annotation based on that, we are expanding our multimodal annotation—for the tiers described in this section—to include the whole length of shows (23 26-min shows).

Our annotation scheme is the result of multiple iterations, careful considerations, and discussions between the members of our multidisciplinary research team (for more details of our work at earlier stages, see Uhrig et al., 2023).

#### 3.1 Textual modality

The textual modality as presented here is an artefact that we include for convenience, fully aware that it is in fact part of the speech signal, which we record on the acoustic channel. From the acoustic channel, Automatic Speech Recognition (ASR) attempts to recognise words for the full files. For the smaller, manually annotated sections, a manual transcription was created by the annotators themselves. Note that any segmentations, e.g., the introduction of punctuation marks in the transcripts, are already interpretations.

<sup>6</sup> <https://beta.rapidannotator.org>

These can be done by a machine in the case of the automatic transcription, where we used automatic punctuation restoration in the preparation of the files for CQPweb (see Uhrig et al., 2023 and Dykes et al., 2023 for details), i.e., the punctuation marks are purely based on the derived textual modality. For the manual transcription, any punctuation marks would also be inspired by prosodic features such as pauses and intonation.

### 3.1.1 Transcript

YouTube provided automatic transcriptions for the videos in our dataset (see Dykes et al., 2023), which are roughly time-aligned on the word level. We import these into ELAN automatically. The manually annotated sections contain a manual transcript, which is, however, not time-aligned on the word level.

There are limitations to this approach in that the word recognition is not always accurate (and the show host's foreign accent slightly reduces the accuracy), so we manually correct the annotations as we proceed with our annotation for individual intervals, although not systematically for entire files.

Furthermore, we have tried a more recent development, Whisper,<sup>7</sup> which on average offers better speech recognition but at the cost of over-standardising (e.g., it removes false starts and hesitation phenomena). Whisper only provides timestamps on the level of an entire subtitle line and not per word, at least not out of the box. For now, we have not pursued this avenue of automatic transcription any further.

### 3.1.2 Classes of future markers

Once we had determined video intervals for viewpoint future depictions, we analysed them for further markers of the future in speech. The analysis allowed us to identify seven classes of future markers in English speech:

- 1 *will*-future
- 2 Conditional clauses and counterfactuals (e.g., *if*-, *when*-clauses)
- 3 Modal verbs (e.g., *should*, *must*)
- 4 Time adverbials (e.g., *in the future*, *next year*)
- 5 *going to*-future and present-tense simple and progressive used with future reference
- 6 Words with a semantic component of future (e.g., *possibility*, *futurist*)
- 7 Words that acquired future semantics within the specific context (e.g., *architect* is defined by the speaker both as an engineer and a futurist, and thus acquires the 'futurist' semantics for the subsequent discourse)

We then proceeded to include in our annotation scheme the tiers for (i) automatic transcription, (ii) viewpoint future depiction; (iii) future marker in speech; and (iv) future marker class.

## 3.2 Acoustic modality

As illustrated in Section 2, integration of all three modalities is important for the analysis of persuasive and manipulative

communication. Thus, it was necessary to identify the boundaries of the principal constituents of the prosodic hierarchy and prominences within these, which can be thought of as prosodic landmarks. By annotating these, we can then proceed to identify whether and how gestural and corporal landmarks align with them.

### 3.2.1 Manual prosodic annotation

Prosodic annotation was done by one or two expert coders manually in Praat<sup>8</sup> and then verified by one to two senior experts before being transferred to ELAN. As showcased in Section 2, the relationship between prosodic structure (e.g., edges of prosodic constituents such as prosodic phrases, prominences within a prosodic constituent) and the acoustic cues to prosody (e.g., pauses, variation in *f0*, duration, and voice quality) is a complex one, with a many-to-one and one-to-many mapping between acoustic cues and prosodic structure. This means that selecting just one acoustic parameter would give not just a partial picture but also one that is also inconsistent in what it depicts. We start with the manual analysis and annotation for prosody—all relevant cues—with the aim of exploring phonetic complexities and laying the groundwork for our future study on the automation of annotation for prosody.

The manual annotation scheme provided below is sufficient to identify two levels of prosodic phrasing (IP and intermediate phrases), prosodic word boundaries, pauses between and within phrases, and two degrees of accentual prominences (phrase accents and the nuclear phrase accent). The annotation was done following the IViE conventions (Grabe et al., 1998). The full process of the manual annotation is described by us in Uhrig et al. (2023: Section 2.7).

### 3.2.2 Manual Annotation Scheme

#### Phrase

**IP** (Intonational Phrase).

**ip** (intermediate phrase).

**ProsWord** (Prosodic Word).

#### Accent

On this tier, all tonal events are labelled:

- 1 pitch accents, which are associated with specific syllables (with specific words), and lend perceptual prominence (the principal one of which in any prosodic phrase is known as the nuclear pitch accent, and marks the prosodic 'head' of that constituent, and the focus of that phrase);
- 2 phrase accents that appear between the last pitch accent and the boundary tone of a phrase;
- 3 boundary tones, which are associated not with words but with the phrase, and appear at the phrase edge, carrying information about the type of phrase (e.g., question vs. statement).

Glossary: L\*, H\*, H\* + L, L\* + H, H-, L-, H%, L%.

#### Nuclear stressed syllable

The nuclear stressed syllable was marked. This aligns with the final pitch accent (i.e., the nuclear pitch accent) on the accent tier.

Glossary: N (Nuclear stress).

<sup>7</sup> <https://openai.com/research/whisper>

<sup>8</sup> <https://www.fon.hum.uva.nl/praat/>

### Comments

On this tier, we noted the following particular prosodic features: mispronunciations, interruptions, speech rate discontinuities, strong focal emphasis, or voice quality effects (Uhrig et al., 2023, Section 2.7).

See the video capturing the annotation for textual and acoustic modalities here.<sup>9</sup>



## 3.3 Visual modality

As far as visual modality is concerned, the case studies on multimodal future depictions by RT that we have done so far have motivated us to annotate gestural movements by hand, face, and head, as well as corporal movements on individual tiers. There is a hierarchical 'annotation' arrangement here since the core focus of our current study is on gestural movements by hand and eyebrows. As explained in sub-section 3.3.2.3, we did not annotate for eye behaviour, gaze movement, head movement, and corporal movement to a full extent since our engagement with those features came secondary out of our primary engagement with the hand and eyebrows.

### 3.3.1 Annotation for gestural units: hand

The complex analysis for meaning, stance, and viewpoint construction that we perform as showcased in Section 2 calls for a fine-grained annotation at a high level of precision.

We therefore started with manual annotation by expert coders for direction and orientation of hand movements. We subsequently worked to automate annotations for hand direction and orientation, guided by both conceptual considerations and constraints posed by the development of the computational tools. We then applied our experience and observations to create a tool for automatic annotation of the direction of hand movements.

We approached our annotation for gestural zones for hands differently, first developing an automatic tool for gestural zone identification and then verifying annotations manually with the help of non-expert coders.

The annotation of gestural zones was more straightforward than the annotation of hand gesture. Our study on algorithms for automatic hand movement detection described in Section 3.3.1.4 allowed for the identification of gestural zones without the need for extensive preliminary manual annotation.

#### 3.3.1.1 Manual hand movement annotation

As demonstrated in Section 2, gestural sequences or individual gestural movements that co-occur with future depictions in speech are complex. To be able to capture the complexity of those on the formal level, we opted for annotating for direction on three axes—sagittal, lateral, and vertical. Having a separate tier for gestural trajectory presented a problem due to the lack of a consistent approach for labelling, which tends to use metaphorical labels and, in doing so, already deviates from the purely formal recording of gestural characteristics. If the gestural sequence or a gestural movement had a complex trajectory, e.g., the 'delete' gestural sequence in our analysis in Section 2, when

the hand goes slightly leftwards but also upwards and then rightwards but also downwards and then just downwards, we captured the complexity by annotating for the same 'gesture' on three tiers—sagittal, lateral, and vertical—for the same interval in speech. Gestural movements recorded as performed along different axes may start and/or end either simultaneously or at different times. Thus, the timings of the sub-intervals created on separate tiers for the same gestural sequence may overlap but do not have to coincide.

That resulted in a situation where we did not have a separate tier for gestural trajectory but still captured the trajectory implicitly across a number of tiers for hand movements. Given that we often encountered gestural movements where the hand, the fingers, and the thumb may be moving, pointing in different directions, or even moving along different axes, we opted for annotating for hands and fingers on separate tiers. For the segmentation of longer gestural sequences into individual units, we relied on two criteria: either a change of direction or a change of axis will delineate individual gestures as we understand them.

This type of annotation took into account the constraints of the potential computer vision tools, for which a small set of categories, e.g., the axes and directions, are easier to distinguish than a complex set of labels.

We annotated for hand and finger movements in a certain direction on six tiers—two for sagittal axis, two for lateral axis, and two for vertical axis—in ELAN with handedness captured through labelling the tiers for axes, e.g., right hand going rightwards would be labelled on the tier for lateral axis as 'RH R'. We had a separate tier for capturing a handshake.

As our approach to analysis and annotation is data-driven, we did not limit ourselves to thinking that future can be conceptualised in gesture for English through forward or rightward movements only. Rather, we used conceptual blending to analyse meaning and viewpoint construction and, through such analysis, to determine whether a certain gestural movement may carry a representational function of future (Section 2). We have analysed examples where we observed various outward-directed hand movements and body-directed hand movements arguably carrying a future function. Therefore, body-directed hand movements were captured as BDG labels on tiers for axes.

The annotation scheme described below was developed to be universally applicable. Although it may not be exhaustive, it has allowed us to capture key parameters of gestural movements with the impressionistically perceived temporal function and to do so through the formal approach to gesture recording.

#### 3.3.1.2 Manual Hand Movement Annotation Scheme

##### Sagittal axis hand

Glossary: RH (Right Hand), LH (Left Hand), BH (Both Hands) + F (Forwards), B (Backwards), BDG B (Body-Directed Gesture Backwards).

##### Lateral axis hand

Glossary: RH (Right Hand), LH (Left Hand), BH (Both Hands) + R (Rightwards), L (Leftwards), S (Spread), C (Centre).

##### Vertical axis hand

Glossary: RH (Right Hand), LH (Left Hand), BH (Both Hands) + U (Upwards), D (Downwards).

<sup>9</sup> See <http://go.redhenlab.org/pgu/0133/> or scan the QR code.





FIGURE 6  
OpenPose keypoints for the screen capture of A43 in Section 2.

### Sagittal axis fingers

Glossary: RH (Right Hand), LH (Left Hand), BH (Both Hands)+ Fingers (all fingers), Thumb, IF (Index Finger), MF (Middle Finger), RF (Ring Finger), LF (Little Finger)+F (Forwards), B (Backwards), and BDG B (Body-Directed Gesture Backwards).

### Lateral axis fingers

Glossary: RH (Right Hand), LH (Left Hand), BH (Both Hands)+ Fingers (all fingers), Thumb, IF (Index Finger), MF (Middle Finger), RF (Ring Finger), LF (Little Finger)+R (Rightwards), L (Leftwards).

### Vertical axis fingers

Glossary: RH (Right Hand), LH (Left Hand), BH (Both Hands)+ Fingers (all fingers), Thumb, IF (Index Finger), MF (Middle Finger), RF (Ring Finger), LF (Little Finger)+U (Upwards), D (Downwards).

### Handshape

The handshape and palm orientation, where applicable, for each moving hand are recorded on a separate tier. The annotation also reflects the changes of handshape during the direction intervals (sagittal/lateral/vertical axis hand).

Glossary: RH (Right Hand), LH (Left Hand), BH (Both Hands)+ OP (Open Palm), CP (Closed Palm), Fist, FB (Finger Bunch), FP (Finger Pinch), Prayer, Cup, Hand brush + A (Away—for OP), B (Back—for OP), U (Up—for OP, Cup), D (Down—for OP, Cup), V (Vertical—for OP).

See the video capturing the manual annotation for hand gesture here.<sup>10</sup>



### 3.3.1.3 Automatic Hand Movement Annotation

The detection of hand gesture is usually done in computer vision by detecting hand movement. Accordingly, most computer vision systems do not distinguish gestures from other types of hand movements, which is in line with our data-driven approach, which rejects the practise of discarding data *a priori*.

Most of our automatic annotation of hand movements is based on body pose estimation, for which we use OpenPose (Cao et al., 2019). With this system, every single frame of the video is annotated with the body keypoints of every person identified.

Figure 6 shows the keypoints for an example from the video analysed in Section 2. Since we rely on media data, the videos do not contain depth information, which means that we only obtain keypoints in 2D space, the *x* and *y* values of which correspond to the vertical and lateral axes as long as the speaker is facing the camera. Since keypoints are detected separately for each frame, we often witness so-called jitter, i.e., small changes of keypoint coordinates between frames without any discernible movement. We use statistical methods to smooth these keypoint positions to eliminate those artefacts introduced by the software, which would otherwise lead to false gesture detections. Furthermore, keypoints may not be detected in some frames, often owing to motion blur. In these cases, we interpolate the coordinates linearly between the last detected and the next detected keypoint, i.e., we draw a straight line between them. If missed detections happen at the beginning or at the end of the scene, we extrapolate the first known or the last known value, respectively, to the beginning and end.

Another problem that the analysis of media broadcasts faces is the frequent changes in camera perspectives, either to give a different perspective of the same person or to switch to showing a different person. Often, both the host and the guest appear next to each other in a split screen. As described in Uhrig et al. (2023: Section 2.3), we automatically cut a video into scenes, deploying active speaker detection and biometric clustering, to obtain annotations for the host of the show only when she is visible and speaking. To account for the differences in speaker size on

<sup>10</sup> See <http://go.redhenlab.org/pgu/0134/> or scan the QR code.

the screen across scenes, we normalised the speaker's size by expressing all hand positions in relation to the average position of the speaker's nose in the scene and normalised to the distance between the average position of the nose and the average position of the neck keypoint. We call this distance our *normalisation unit*.

In a first step of automation, we added time series of the wrist keypoints of both hands for the vertical and lateral axes to ELAN (see short description in Uhrig et al., 2023, Section 2.3). The videos<sup>11</sup> (also taken from Uhrig et al., 2023) show the time series for wrists in the second and third time series panels at the top of ELAN's annotation window. Despite the normalisation procedure outlined above, we can still observe shifts in the time series plot when there is a scene change.

During the manual annotation phases, we established that the wrist keypoints were generally reliable when detected. In order to further speed up and support the manual annotation process, we added a rule-based direction detection on the vertical and lateral axes. Our system detects any movement of the smoothed wrist keypoint (separately for the left and right wrist) that goes in the same direction (i.e., leftwards or rightwards for the lateral axis and upwards or downwards for the vertical axis) for at least six frames (i.e., 0.24 s).

The system is highly sensitive to even very small movements that are hardly visible to the naked eye and may well be just artefacts of the computer vision system's calculations. We introduced a threshold below which we do not detect, corresponding to roughly 1 mm difference per frame, in order to reduce the number of these wrongly detected 'gestures'. The exact value of this threshold is currently being evaluated in close conjunction with the manual annotation experts. Therefore, both the unfiltered and the filtered versions of the tier exist in parallel in our ELAN files.

While the system is reliable for most of the data, there are limitations with respect to certain camera perspectives. For instance, at the end of the video snippet above, the speaker is filmed diagonally from behind, sitting in front of a large screen. Here, the direction information is lacking, also because often the right hand is occluded by the body of the speaker. Another problematic case is illustrated by the video snippet analysed in Section 2 above, where in the close-up shots, only the hands are visible from time to time but never the elbow of the speaker. In such cases, OpenPose cannot detect the wrist as part of the speaker's body because the connection via the elbow keypoint is missing. If this happens for the entire scene, even the interpolation method outlined above cannot help because there are not enough data points available. We are currently evaluating the use of other pose estimation systems that detect hands separately, even if the elbow is not detected.

As demonstrated in Section 2, the Speaker's hand position in relation to her body is important in the analysis of time conceptualisation in gesture—e.g. if a hand movement with a future function is made within the central gestural zone, that may signal that the Speaker does not believe that the future event depicted will materialise. Because our data are 2D, we have so far automatically annotated for the vertical and lateral axes only. From a conceptual perspective, we have adapted to the needs of our analysis of McNeill's gesture space diagram (1992: 89). In our adapted diagram (see Figure 7), we distinguish 17 zones. These zones are a combination of boundaries along the vertical and lateral axes.

To automatically identify those 17 gestural zones, we follow the approach described in Section 3.3.1.3, i.e., we make use of normalised, smoothed, and interpolated keypoint coordinates with reference points and normalisation units. We start out by working with both axes separately and identifying five different zones for each. Different reference points and normalisation units are defined for each axis. For the vertical axis, the reference point is a nose  $y$ -coordinate, and the normalisation unit (NU) is the distance between nose and neck, as mentioned in Section 3.3.1.3. We match the vertical position of the wrists to the zones defined in Figure 7, e.g., if the wrist's  $y$ -coordinate is below the reference point by more than three times the length of our normalisation unit, we assign the "Down" label to it. The full list of criteria is given in Table 2.

For the lateral axis, the reference point is the neck  $x$ -coordinate. We use different normalisation units for the right and left wrists. The normalisation unit for the right wrist (RNU) is the horizontal distance between the neck  $x$ -coordinate and the  $x$ -coordinate of the right shoulder, and the normalisation unit for the left wrist (LNU) is the horizontal distance between the neck  $x$ -coordinate and the  $x$ -coordinate of the left shoulder, respectively. Although we generally observe similar values for RNU and LNU, having two independent reference units minimises the effect of jitter discussed in Section 3.3.1.3 and thus leads to more consistent results.

For the horizontal position of the wrists defined in Figure 7, we use the criteria given in Table 3.

As a result, for each frame, we obtain a pair of labels for a wrist position (vertical position label and horizontal position label). We then merge and rename these to produce final labels in accordance with the predefined gestural zones in Figure 7.

#### Time series panels:

- Right Wrist Lateral Position, Left Wrist Lateral Position
- Right Wrist Vertical Position, Left Wrist Vertical Position

#### Tiers:

- Right Wrist Lateral Direction Auto  
Glossary: Right, Left
- Right Wrist Lateral Direction Auto (Threshold)  
Glossary: Right, Left
- Right Wrist Vertical Direction Auto  
Glossary: Up, Down
- Right Wrist Vertical Direction Auto (Threshold)  
Glossary: Up, Down
- Left Wrist Lateral Direction Auto  
Glossary: Right, Left
- Left Wrist Lateral Direction Auto (Threshold)  
Glossary: Right, Left
- Left Wrist Vertical Direction Auto  
Glossary: Up, Down
- Left Wrist Vertical Direction Auto (Threshold)  
Glossary: Up, Down
- Right Wrist Zone Auto

Glossary: Right Up, Up, Left Up, Centre Right Up, Centre Up, Centre Left Up, Right, Centre Right, Centre, Centre Left, Left, Right Down, Centre Right Down, Centre Down, Down, Centre Left Down, Left Down

<sup>11</sup> <http://go.redhenlab.org/pgu/0130> and <http://go.redhenlab.org/pgu/0131>

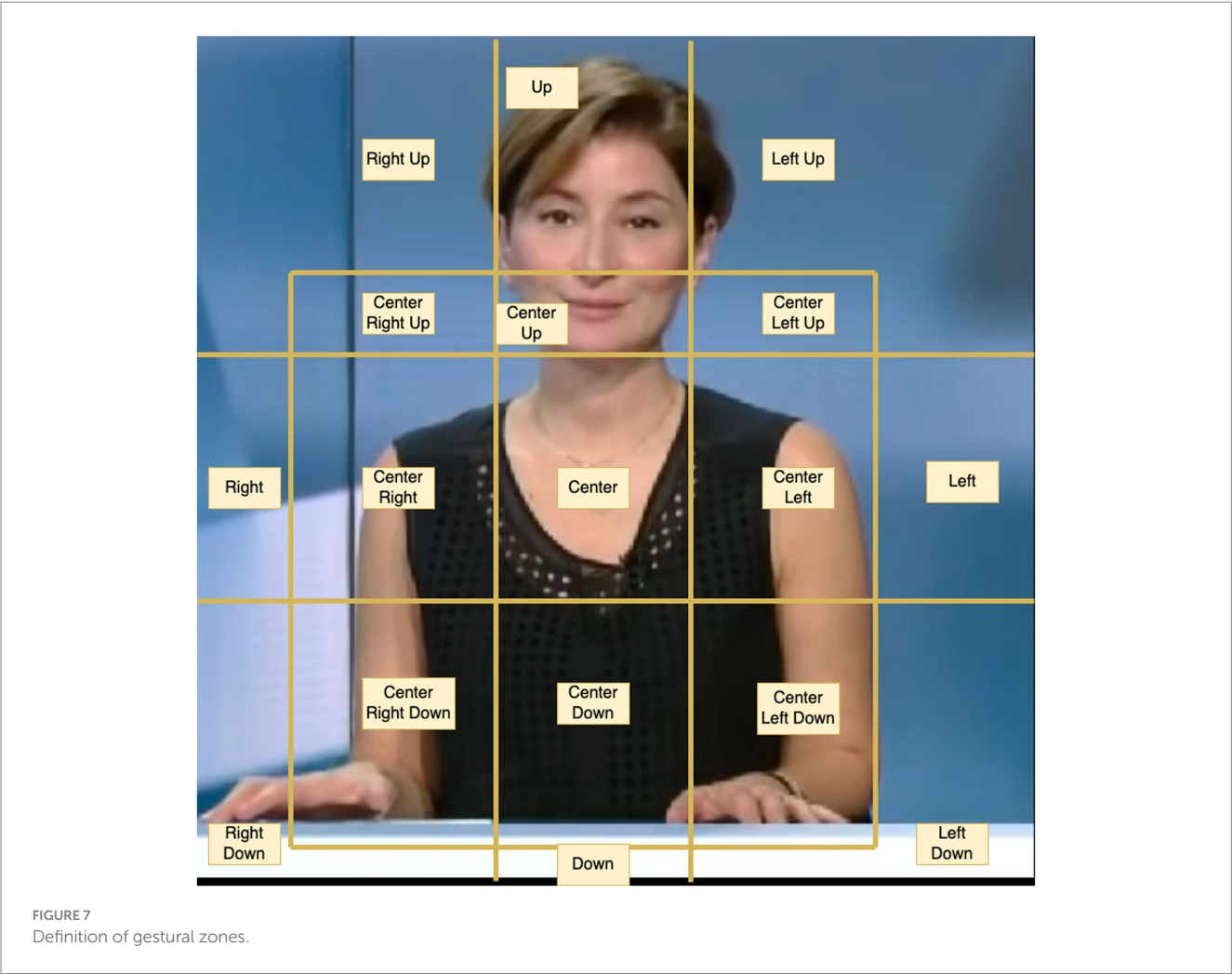


TABLE 2 Criteria for distinguishing gestural zones on the vertical axis.

Normalised y-coordinate value	Label
$y < -3.5 \times \text{NU}$	Down
$-3.5 \times \text{NU} \leq y < -2 \times \text{NU}$	Centre down
$-2 \times \text{NU} \leq y < -0.5 \times \text{NU}$	Centre
$-0.5 \times \text{NU} \leq y < 0$	Centre up
$y \geq 0$	Up

• Left Wrist Zone Auto

Glossary: Right Up, Up, Left Up, Centre Right Up, Centre Up, Centre Left Up, Right, Centre Right, Centre, Centre Left, Left, Right Down, Centre Right Down, Centre Down, Down, Centre Left Down, Left Down.

See the video capturing the automatic annotation for hand movement and gestural zones here <sup>12</sup>.



<sup>12</sup> See <http://go.redhenlab.org/pgu/0135> or scan the QR code.

TABLE 3 Criteria for distinguishing gestural zones on the lateral axis.

Normalised x-coordinate value	Label
$x < -1.5 \times \text{RNU}$	Right
$-1.5 \times \text{RNU} \leq x < -0.75 \times \text{RNU}$	Centre right
$-0.75 \times \text{RNU} \leq x < 0.75 \times \text{LNU}$	Centre
$0.75 \times \text{LNU} \leq x < 1.5 \times \text{LNU}$	Centre left
$x \geq 1.5 \times \text{LNU}$	Left

3.3.2 Annotation for gestural units: eyebrows

We manually annotated several video clips for facial gesticulation at the initial exploratory stage. Since facial gesticulation cannot be directly linked to temporal representation, we opted to annotate for facial gesticulation throughout videos and did not restrict it to specific temporal speech-led intervals. This approach proved to be too time-consuming and labour-intensive and could not be sustained. At the same time, the exploratory annotation for facial gesticulation informed by our studies allowed us to make a better-informed choice as to what facial gestural feature to annotate first for the purposes of our study on future depictions. We chose to annotate for eyebrow movement along the vertical axis. As showcased in Section 2, eyebrow movements are

coordinated with other types of gestural movements (e.g., hand) and with prosodic organisation in meaning and stance construction. For example, eyebrow movements, like prosody, mark prominences and phrase boundaries, contributing to the construction of sceptical stance.

Furthermore, it proved possible to develop an automatic annotation tool for eyebrow movements quickly. This enabled automatic processing of our data for eyebrow movement visualisation first. We then proceeded to analyse automatic eyebrow visualisation time series in an exploratory fashion. We developed an approach for subsequent manual verification of automatic annotation for eyebrows and the addition of further manual annotation. Not only did such an approach allow us to considerably speed up annotation, but it also enabled us to gather further insights into coordination between eyebrow movements, eye behaviour, and head gesticulation that we would not have spotted otherwise. The computer vision algorithm highlighted for us small movements, which we would have ignored during the fully manual annotation due to the richness of the data and the limitations of human attention.

### 3.3.2.1 Automatic annotation of eyebrow movement

The automatic eyebrow visualisation time series indicates the vertical position of each eyebrow. We cannot, however, equate this to the eyebrow's vertical position in the video frame because head movements (and particularly head tilts) adversely affect the calculation of the position in relation to a facial landmark. In our software, we use OpenPose's face keypoints with the same kind of smoothing and interpolation as described above for the hand gestures. The normalisation unit is the distance between the top and the tip of the nose. We calculate the mean position of the eyebrow keypoints and compare this to the mean position of the lower eyelid keypoints. We inherit certain issues from the limitations of OpenPose. Thus, during blinks or longer periods of closed eyes, the keypoints of the lower lid are detected further down, which leads to a relatively higher position of the eyebrows. At first, we regarded this as a flaw in the automatic visualisation, but upon further inspection, we decided that even these blinks and closed eyes may be meaningful units for our analysis of facial gesticulation and its role in the overall meaning, viewpoint, and stance construction, as illustrated in Section 2. We do not know how and in relation to precisely which facial movements humans perceive raised eyebrows, so these cases might function perceptually in a manner similar to raised eyebrows. As explained in further detail in the next sub-section, we opted to do manual annotation for eye, head, and corporal behaviour only as prompted by eyebrow movement, or what the machine, in contrast to human coders, saw as eyebrow movement. Our annotation for facial and head gesticulation or eye or corporal behaviour is by no means exhaustive. It serves the purpose of our ongoing analysis of multimodal depictions of futures as showcased in Section 2 and is, at this stage, exploratory.

Time series panel:

- Right Eyebrow Vertical Position, Left Eyebrow Vertical Position

### 3.3.2.2 Validation and manual annotation of eyebrow movement and related phenomena

The automatic tracking of eyebrow movement was reliable in most cases but still had some limitations due to such factors as scene change, head movement, and poor video quality.

Two coders went through all eyebrows time series to establish whether the OpenPose-based movement detection was correct. They created corresponding intervals to note the direction of the eyebrow movement and establish the boundaries of the eyebrow units. Disagreements regarding the boundaries were resolved through discussion.

Coders were observed and annotated for two kinds of errors in automatic annotation. The first was when the machine produced an error that could not be explained by what human coders saw in the video, e.g., the Speaker's gesticulation, corporal behaviour, or hair masking the eyebrows. The second kind of error could be explained by factors such as scene change, head movement, and poor video quality that the human coders encountered.

### 3.3.2.3 Manual annotation scheme

#### Eyebrow movements

Glossary: BU (Both eyebrows Up), BD (Both eyebrows Down), LU (Left eyebrow Up), LD (Left eyebrow Down), RU (Right eyebrow Up), RD (Right eyebrow Down).

#### Peak or plateau

We differentiated between two types of eyebrow movement: Peak and Plateau. These are working terms emerging from our exploratory analysis that are not grounded in any theoretical framework offered elsewhere. As we proceed with our analysis, we may opt to change the terms and/or offer a new theoretical framework emerging from our observations and analysis.

In our engagement with Peak and Plateau as working terms and concepts, we relied on the length of the domain, which coincides with eyebrow movement, as the criterion. We defined Peak as a short accent-like eyebrow movement (its domain can be a word, prosodic word, or a syllable) and Plateau as a prolonged movement where eyebrows would stay in the same position for a longer time (its domain can be an ip, IP, grammatical clause, or a sentence/phrase).

Glossary: Pk (Peak), Pl (Plateau).

#### Head movement

Glossary: TL (Tilt Left), TR (Tilt Right), TD (Tilt Down), TU (Tilt Up), TF (Tilt Forward), TB (Tilt Backward), Tr L (Turn Left), Tr R (Turn Right), Nod.

When coders impressionistically perceived a head tilt forward as a nod, they recorded it as such, but if in any doubt whatsoever, they annotated it as 'head tilt forward'. We included the term *tilt* deliberately to avoid using a more *loaded*—linked to a function—label and annotated smooth (impressionistically) head movements as tilts in an attempt to capture only the formal side of head gestural movements.

#### Corporal movement

Glossary: F (Forward), B (Backward), U (Upward), D (Downward), Shrug (Shoulder shrug).

#### Eyes

Glossary: O (Open wide), B (Blinking), S (Squinting), Cl (Closed eyes), W (Winking).



### Gaze

Glossary: U (Up), D (Down), L (Left), R (Right), UL (Up Left), UR (Up Right), DL (Down Left), DR (Down Right).

### Gaze focus

Glossary: F (Focused), D (Distanced).

See the video capturing the automatic annotation for eyebrow movement and manual annotation for gesticulation and corporal behaviour here.<sup>13</sup>



## 4 Conclusion

We have presented our annotated multimodal dataset and the methodology underpinning its creation. Our case study showcased the necessity of including in our annotation scheme various tiers and features from three modalities: textual, acoustic, and visual.

The implementation of our approach has relied on ongoing dialogue between our team's linguists (experts in cognitive linguistics, discourse analysis, phonetics and prosody, gesture study, computational analysis, and area studies), engineers, and computer scientists (cf. Bateman, 2022a, p. 59). Through this interdisciplinary work, we have been able to produce methodologically sound analyses and computationally tractable annotations. We have extended the framework of conceptual integrating/blending as a cognitive theory to explore how cues of speech (including prosody), gesticulation, and corporal behaviour work together to construct meaning, stance, and viewpoint in RT communication and translate our insights into decisions on annotation strategies. Our automatic annotations used theoretically informed categories, and our manual annotations were adjusted for optimal use in machine learning.

Our study continues, and, among other things, we are producing automatic annotation for amplitude and velocity of gestural movements, which our case studies have shown to be important to include in our dataset.

We envisage using our annotated dataset not just for the purposes of generalising our ongoing multimodal analysis of RT's depictions of the future but also for fine-tuning a multimodal model pre-trained on big data from RT using unsupervised machine learning. To this end, we have already begun to leverage more advanced AI methods to the benefit of all disciplines involved in our multimodal research.

On the conceptual side, our ability to identify the relevant variables within each modality at scale and speed and to see patterns now opens a pathway for building a new theoretical model for speech–gesture interaction.

## Primary sources

RT show 'SophieCo Visionaries', episode 'We've lost control of our phones', downloaded from YouTube, last accessed on 3 February 2022 (<http://go.redhenlab.org/pgu/0138>).

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material; further inquiries can be directed to the corresponding author.

## Ethics statement

Ethical approval was not required for the study involving human data in accordance with the local legislation and institutional requirements. Written informed consent was not required for participation in study or for the publication of identifying images or data in accordance with the local legislation and institutional requirements.

## Author contributions

AW: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualisation, Writing – original draft, Writing – review & editing. IP: Writing – original draft, Writing – review & editing, Data curation, Investigation, Validation, Visualization. EP: Conceptualization, Formal analysis, Methodology, Supervision, Writing – review & editing. IB: Methodology, Software, Writing – original draft. PU: Data curation, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The research presented in this article was made possible by generous funding provided by the Arts and Humanities Research Council (grant reference AH/W010720/1), the Deutsche Forschungsgemeinschaft (project number 468466485), and The John Fell Oxford University Press Research Fund (grant reference number 0011177). The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Centre (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project b105dc. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG)—440719683.

## Acknowledgments

The authors would like to thank: Georgy Krasovitsky for help with gesture annotation; Anna Sapuntsova for help with annotation for prosody; Evie Burrows for annotation for prosody and eyebrow movements; Mary Baltazani for expert engagement with prosody annotation; Scott Hale for help with collecting and storing data; Nathan Dykes for corpus building and NLP; Philip Torr and

<sup>13</sup> See <http://go.redhenlab.org/pgu/0136> or scan the QR code.

N. Siddharth for advising on computer vision-related work; and Andrew Wilson for proofreading the manuscript.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

- Andries, F., Meissl, K., de Vries, C., Feyaerts, K., Oben, B., Sambre, P., et al. (2023). Multimodal stance-taking in interaction—A systematic literature review. *Front. Commun.* 8:1187977. doi: 10.3389/fcomm.2023.1187977
- Bakhtin, M. (2013). *Problems of Dostoevsky's Poetics*. United States: University of Minnesota Press.
- Bateman, J. A. (2022a). Multimodality, where next? – some meta-methodological considerations. *Multimod. Soc.* 2, 41–63. doi: 10.1177/26349795211073043
- Bateman, J. A. (2022b). Growing theory for practice: empirical multimodality beyond the case study. *Multimod. Commun.* 11, 63–74. doi: 10.1515/mc-2021-0006
- Bavelas, J. B., Chovil, N., Lawrie, D. A. (1992). Interactive gestures. *Discourse Processes*. 15, 469–489. doi: 10.1080/01638539209544823
- Brône, G., Oben, B., Jehoul, A., Vranjes, J., and Feyaerts, K. (2017). Eye gaze and viewpoint in multimodal interaction management. *Cogn. Linguist.* 28, 449–483. doi: 10.1515/cog-2016-0119
- Cao, Z., Gines, H., Tomas, S., Shih-En, W., and Yaser, S. (2019). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 172–186. doi: 10.48550/arXiv.1611.08050
- Chu, M., Meyer, A., Foulkes, L., and Kita, S. (2014). Individual differences in frequency and saliency of speech-accompanying gestures: the role of cognitive abilities and empathy. *J. Exp. Psychol. Gen.* 143, 694–709. doi: 10.1037/a0036311
- Cooperrider, K., Núñez, R., and Sweetser, E. (2014). “The conceptualization of time in gesture. Body-language-communication” in *Body-Language-Communication: An International Handbook on Multimodality in Human Interaction*. eds. C. Müller, A. Cienki, E. Fricke, S. Ladewig, A. McNeill and J. Bressems, vol. 2 (Berlin, München, Boston: De Gruyter Mouton), 1781–1788.
- Dancygier, B., Hinnell, J., and Lou, A. (2019). “Stance construction in multimodal, multi-media contexts” in *Paper presented at the 15th International Cognitive Linguistics Conference*. Nishinomya, Japan
- Dancygier, B., and Sweetser, E. (2000). Constructions with if, since, and because: causality, epistemic stance, and clause order. *Top. English Linguist.* 33, 111–142.
- Dancygier, B., and Sweetser, E. (2005). *Mental Spaces in Grammar: Conditional Constructions*. New York: Cambridge University Press.
- Dykes, N., Wilson, A., and Uhrig, P. (2023). “A pipeline for the creation of multimodal corpora from YouTube videos” in *Proceedings of Linguistic Insights from and for Multimodal Language Processing (LIMO 2023)* at KONVENS, Ingolstadt.
- Eijk, L., Rasenberg, M., Arnese, F., Blokpoel, M., Dingemanse, M., Doeller, C. F., et al. (2022). The CABB dataset: A multimodal corpus of communicative interactions for behavioural and neural analyses. *NeuroImage* 264:119734. doi: 10.1016/j.neuroimage.2022.119734
- Fauconnier, G. (1994). *Mental Spaces: Aspects of Meaning Construction in Natural Language*. Cambridge: Cambridge University Press.
- Fauconnier, G., and Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.
- Grabe, E., Nolan, F., and Farrar, K. (1998). “IViE—A comparative transcription system for intonational variation in English” in *Proceedings of ICSLP 98*, Sydney, Australia.
- Hayes, B. (1989). “The prosodic hierarchy in meter” in *Phonetics and Phonology*. eds. P. Kiparsky and G. Youmans, vol. 1 (San Diego: Academic Press), 201–260.
- Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychol. Bull.* 137, 297–315. doi: 10.1037/a0022128
- Jowett, G., and O'Donnell, V. (2006). *Propaganda and Persuasion*. Thousand Oaks/London/New Delhi: Sage Publications.
- Kibrik, A. (2018). Russian multichannel discourse. Part II. Corpus development and avenues of research [Russkiy mul'tikanal'nyy diskurs. Chast' II. Razrabotka korpusa i napravleniya issledovaniy]. *Psikh. Zhurnal* 39, 78–89. doi: 10.7868/80205959218020083
- Kita, S. (2000). “How representational gestures help speaking” in *Language and Gesture*. ed. D. McNeill (Cambridge: Cambridge University Press), 162–185.
- Kok, K., Bergmann, K., Cienki, A., and Kopp, S. (2016). Mapping out the multifunctionality of speakers' gestures. *Gesture* 15, 37–59. doi: 10.1075/gest.15.1.02kok
- Loehr, D. (2014). “Gesture and prosody” in *Body-Language-Communication: An International Handbook on Multimodality in Human Interaction*. eds. C. Müller, A. Cienki, E. Fricke, S. Ladewig, A. McNeill and J. Bressems, vol. 2 (Berlin, München, Boston: De Gruyter Mouton), 1381–1391.
- Lücking, A., Bergmann, K., Hahn, F., Kopp, S., and Rieser, H. (2010). “The Bielefeld speech and gesture alignment Corpus (SaGA)” in *LREC 2010 workshop: Multimodal corpora—advances in capturing, coding and analyzing multimodality*. (eds.) M. Kipp, M.J.-P. Martin, P. Paggio, and D. Heylen, 92–98.
- McNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. Chicago: University of Chicago Press.
- Mittelberg, I. (2018). Gestures as image schemas and force gestalts: A dynamic systems approach augmented with motion-capture data analyses. *Cogn. Semiot.* 11:1. doi: 10.1515/cogsem-2018-0002
- Narayan, S. (2012). “Maybe what it means is he actually got the spot: physical and cognitive viewpoint in a gesture study” in *Viewpoint in Language: A Multimodal Perspective*. eds. B. Dancygier and E. Sweetser (Cambridge: Cambridge University Press), 97–112.
- Nespor, M., and Vogel, I. (1986). *Prosodic Phonology*. Dordrecht: Foris Publications.
- Núñez, R., and Cooperrider, K. (2013). The tangle of space and time in human cognition. *Trends Cogn. Sci.* 17, 220–229. doi: 10.1016/j.tics.2013.03.008
- Parrill, F. (2012). “Interactions between discourse status and viewpoint in co-speech gesture” in *Viewpoint in Language: A Multimodal Perspective*. eds. B. Dancygier and E. Sweetser (Cambridge: Cambridge University Press), 97–112. doi: 10.1017/CBO9781139084727.008
- Parrill, F., and Sweetser, E. (2004). What we mean by meaning: conceptual integration in gesture analysis and transcription. *Gesture* 4, 197–219. doi: 10.1075/gest.4.2.05par
- Pascual, E. (2014). *Fictive Interaction: The Conversation Frame in Thought, Language, and Discourse*. Amsterdam-Philadelphia: John Benjamins Publishing Company.
- Pierrehumbert, J. B. (1980). The phonology and phonetics of English intonation. PhD Dissertation. MIT.
- Pleshakova, A. (2018). “Cognitive approaches: media, mind, and culture” in *The Routledge Handbook on Language and Media*. eds. C. Cotter and D. Perrin (London and New York: Routledge), 77–93.
- Pouw, W., Trujillo, J., Bosker, H. R., Drijvers, L., Hoetjes, M., Holler, J., et al. (2023). *Gesture and Speech in Interaction (GeSpIn) Conference*. doi: 10.17617/2.3527196
- Rühlemann, C., and Ptak, A. (2023). Reaching beneath the tip of the iceberg: A guide to the Freiburg multimodal interaction Corpus. *Open Linguist.* 9:1. doi: 10.1515/opli-2022-0245
- Selkirk, E. (2003). “Sentence phonology” in *The Oxford International Encyclopedia of Linguistics*. eds. W. Frawley and W. Bright. 2nd ed (New York and Oxford: Oxford University Press).
- Sinclair, J. M. C. H. (1991). *Corpus Concordance Collocation*. Oxford: Oxford University Press.
- Steen, F., Hougaard, A., Joo, J., Olza, I., Cánovas, C. P., Pleshakova, A., et al. (2018). Toward an infrastructure for data-driven multimodal communication research. *Linguist. Vanguard.* 4:1. doi: 10.1515/lingvan-2017-0041
- Sweetser, E. (2012). “Introduction: viewpoint and perspective in language and gesture, from the ground down” in *Viewpoint in Language: A Multimodal Perspective*. eds. B. Dancygier and E. Sweetser (Cambridge: Cambridge University Press), 1–22.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Tobin, V. (2017). Viewpoint, misdirection, and sound design in film: *the conversation*. *J. Pragmat.* 122, 24–34. doi: 10.1016/j.pragma.2017.06.003
- Turner, M. (2014). *The Origin of Ideas. Blending, Creativity, and the Human Spark*. New York: Oxford University Press.
- Turner, M., Avelar, M., and Mendes de Oliveira, M. (2019). Atenção Compartilhada Clássica Mesclada e Dêixis Multimodal. *Signo* 44, 3–9. doi: 10.17058/signo.v44i79.12710
- Uhrig, P., Payne, E., Pavlova, I., Burenko, I., Dykes, N., Baltazani, M., et al. (2023). “Studying time conceptualisation via speech, prosody, and hand gesture: interweaving manual and computational methods of analysis” in *Gesture and Speech in Interaction (GeSpIn) Conference*. (eds.) W. Pouw, J. Trujillo, H. R. Bosker, L. Drijvers, M. Hoetjes, J. Holler, L. Van Maastricht, E. Mamus, and A. Ozyurek.
- Ungerer, F., and Schmid, H.-J. (2013). *An Introduction to Cognitive Linguistics*. London, New York: Routledge.
- Valenzuela, J., Pagán Cánovas, C., Inés, O., and Carrión, D. A. (2020). Gesturing in the wild: evidence for a flexible mental timeline. *Rev. Cogn. Linguist.* 18, 289–315. doi: 10.1075/rcl.00061.val
- Vandelanotte, L. (2017). “Viewpoint” in *The Cambridge Handbook of Cognitive Linguistics*. Ed. B. Dancygier (Cambridge: Cambridge University Press), 2, 157–171.
- Wilson, A., Wilkes, S., Teramoto, Y., and Hale, S. (2023). Multimodal analysis of disinformation and misinformation. *R. Soc. Open Sci.* 10:230964. doi: 10.1098/rsos.230964
- Xiang, M., and Pascual, E. (2016). Debate with Zhuangzi: expository questions as fictive interaction blends in ancient Chinese philosophy. *Pragmatics* 26, 137–162. doi: 10.1075/prag.26.1.07xia
- Yoon, Y., Ko, W.-R., Jang, M., Lee, J., Kim, J., and Lee, G. (2019). “Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots” in *Proceedings of the International Conference in Robotics and Automation (ICRA)*.



## OPEN ACCESS

## EDITED BY

Janina Wildfeuer,  
University of Groningen, Netherlands

## REVIEWED BY

Loli Kim,  
University of Oxford, United Kingdom  
Chiao-I Tseng,  
University of Gothenburg, Sweden

## \*CORRESPONDENCE

Karl-Heinrich Schmidt  
✉ khschmi@uni-wuppertal.de

RECEIVED 15 November 2023

ACCEPTED 10 April 2024

PUBLISHED 01 May 2024

## CITATION

Schmidt K-H (2024) How films convey meaning through alternating structures (with an illustrative analysis of *The Sunbeam*). *Front. Commun.* 9:1338813. doi: 10.3389/fcomm.2024.1338813

## COPYRIGHT

© 2024 Schmidt. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# How films convey meaning through alternating structures (with an illustrative analysis of *The Sunbeam*)

Karl-Heinrich Schmidt\*

School of Electrical, Information and Media Engineering, University of Wuppertal, Wuppertal, Germany

Films and texts differ in terms of their possible logical structures and freedom of presentation on an output medium. While texts can be structured at any depth, the capabilities for structuring films are generally limited. In the presentation of textual documents, the sentence order is usually preserved, whereas video documents often allow rearrangements that lead to new alternations of shots. The fundamental difference between textual and video structures is taken as a starting point. Then, based on a detailed analysis of two different layouts of the film *The Sunbeam* by D. W. Griffith, a formal criterion for distinguishing between internal, discursively motivated and external, diegetically motivated alternations is developed. The results enable a new approach to alternation in film analysis and production.

## KEYWORDS

media theory, document theory, film analysis, editing, alternation, crosscutting, Griffith, *The Sunbeam*

## 1 Introduction

Moving image data can be realized in books (e.g., flip books, to take an example from media history), and texts can be integrated into films (with the potential to have a significant effect, even outside the opening and closing credits, in the age of silent film in particular; see below)—but it is only with media convergence as manifested on screens through the World Wide Web that it becomes possible to employ them with equal weight. This is a significant motivation for treating them together with identical basic concepts. The following analysis thus draws on digital document processing, which in recent decades has been refined specifically for multimedia documents and, given the prevalence of electronic documents, is now ubiquitous. In addition to texts, moving image data is increasingly prominent among the content architectures used in multimedia documents. Similar to texts, these are easily structured and may be processed in a grammar-oriented manner. Therefore, the following section, **Three perspectives on documents**, introduces a general scheme for electronic documents. This approach can also be understood as a contribution to a common metalanguage for multimodal corpora.

In the section **Structures in (video) documents**, we initially focus on the logical structure of text and video documents, which already show fundamental differences. Then, we investigate the standard layout principle of alternation for video documents. Speaking generally, a structured video document involves a display of different shots (such as playing cards). If a video document is played back in sequence in an output stream, some cards in the display can



be switched around locally without adversely affecting the whole document; others should remain in order locally, as their sequence is significant. Identifying such orders also makes it possible to distinguish between internal (chosen within a given discourse) and external (grounded in diegesis) forms of alternation.

The motion picture *The Sunbeam* and a prominent remake of that movie will be taken as an empirical basis for the analysis of alternations in video documents. They will be introduced in the section *The Sunbeam* and classified according to their phenotypes at the level of the shot, that is, the visual appearance of individual shots. No analysis of the underlying binary data is used additionally.

Using the theoretical tools provided earlier and the example, an empirical analysis of the structure of a story then follows in the section **Constructing the story**. In the empirical analysis, the diegetic space will first be constituted for the entirety of the shots by mapping spatiotemporal events; building on this, the continuous space–time regions represented in scenes will be labeled and the diegetic progression deconstructed into sequences. This yields a basal logical structure of the example document, for which the framework of the story will then be identified.

The section **Progressive spatial transitions and alternations** will be followed by an evaluation. To that end, so-called progressive spatial transitions and progression bridges will be identified in a micro-analysis of video segments. These are critical points of understanding for a viewer and provide the basis for a general criterion for distinguishing between internal and external alternations.

In the section **Discussion**, the results are linked back to the current view of document processing with multimodal content.

## 2 Three perspectives on documents

The constructs that we apply are drawn from a document framework that naturally includes a far broader range of artifacts than films as traditionally conceived in film studies—this locates film, including both narrative and documentary films, against the broader background of video surveillance, video protocols of medical operations, visually displayed temporally dependent information, interactive animations, and many more. For documents in general, one can essentially adopt three perspectives: the content view, the logical view, and the layout view.<sup>1</sup>

The content view perspective covers the typical user interest in a document: that is, assuming for now a range of intended readers, viewers, or hearers, what these will generally orient toward will be the represented content of the document. Although much can be said about such content, we will only consider this view to the extent that it is relevant for building our analytic framework. From the document perspective, the notion of content used corresponds to the body of material that has, by some means, been selected for presentation within some document; with respect to the document, therefore, it can be seen as pre-existing, and the main question concerns the organization that is imposed upon it to construct a document. Content portions can employ various content architectures—for output on a

flat surface, this often means plain text, images, or moving images whose data formats can be identified, for instance, by their own MIME types.<sup>2</sup>

Since we will be focusing exclusively on video documents, the content will be taken as raw recordings or creations of some pro-filmic material. This can be taken as corresponding loosely to the various takes produced during filming before being edited into their appearance in the final video document. Therefore, the shot serves as a typical example of a content portion. We will also assume that this content can, at least in principle, be labeled with respect to some place of occurrence and a time of occurrence. This content, therefore, makes available particular space–time slices of some real or created world.

For material to become useful as a document, it is necessary to provide its users (either human or machine) with a way of structuring its content. This is achieved first by imposing a logical organization on that content. This logical view, in essence, covers part-whole relationships, groups content portions into larger structures of related content, and is typically modeled as a tree structure. In a text such as this one, for example, the logical view models such properties as a sentence considered as a part of a paragraph, the paragraph as part of a chapter, and so on. For characterizing the overall structural organization of a document, it is the logical view that is decisive and prior. For the film, the logical organization might then characterize “scenes” as grouped into “acts” and “acts” as making up the entire film (cf. Kavin, 1992, pp. 68–69 with application to the crime film *The Godfather*). In this context, the work of basic film interpretation will come down to reconstructing the logical organization based on the audiovisual material presented to a viewer.

Finally, to make a document readable for humans, there is the further step of selecting a particular layout for the logical organization. This prepares rendering of the content of a document for presentation on some output device or display medium, such as a sheet of paper and a display screen. Thus, any document is seen as a collection of logically organized content that is rendered appropriately for display in some output medium. The actual rendering, i.e., selecting and converting content portions, is where the layout process plays its role. This process is responsible for allocating content to particular forms of presentation and allocating these to, for visual documents, geometrically describable layout objects that can then be displayed on the output medium. Typically, such presentations are also more or less richly structured; we term the result of the layout process as layout structure. Any document artifact is, therefore, to be seen as the result of performing a layout process. This determines the final form of the presentation as accessible to its recipients.

2 The following pages operate primarily with the text and video content architectures. These two content architectures were originally specified in the Multipurpose Internet Mail Extensions (MIME) standard as categories for different media types. See <https://www.iana.org/assignments/media-types/media-types.xhtml> (08.02.2024). A video document consists of content portions of MIME-type video. Every digitized film is a video document if converted to a video MIME-type. We draw a distinction between video as a content architecture and film as a form of presentation (normally with front matter and back matter comparable to that of a book in the guise of opening and closing credits; cf. Schlupkoth and Schmidt, 2022, p. 109).

1 Portions of this and the following section are based on the studies by Schlupkoth and Schmidt (2022) and Bateman and Schmidt (2011).

3 Structures in (video) documents

3.1 Structured documents

Unstructured documents do not provide any viewer-independent specifications for identifying subdocuments (this is the case with many photographs); structured documents do exactly that. For the structured case, both the logical view and the layout view allow decompositions. In both cases, these are seen as hierarchical tree organizations.

In our conceptual framework for structured documents, we use the basic architecture model for document processing in ISO/IEC 8613-2 (1993).

For the logical view of any given structured document, ISO/IEC 8613-2 (1993) defines various types of nodes:

- document logical root.
- composite logical objects.
- basic logical objects.

The document’s logical root is the logical object that is the ancestor of all the other logical objects, and it can contain any number and combination of composite and basic logical objects. A composite logical object is the child of another composite logical object or the document logical root. This, in turn, can contain any number (greater than zero) and a combination of composite or basic logical objects. A basic logical object is a terminal node in the tree structure that can host content portions and does not contain any further logical objects. The structural depth of the logical view of a document is simply the number of levels between the document’s logical root and the basic logical objects. Figure 1 shows, on the left, a possible logical document structure that will be associated with real-world documents in what follows.

A tree structure can be generated analogously for layouts (see Figure 1, middle). Layout structure and logical structure are independent of each other and can, therefore, diverge. However, as illustrated in Figure 1, both share the same content portions, which

are divided between the basic objects of the layout structure in a layout process.

As our first example for discussion, we will take the following five-line text underlying the Christian *ichthus*:

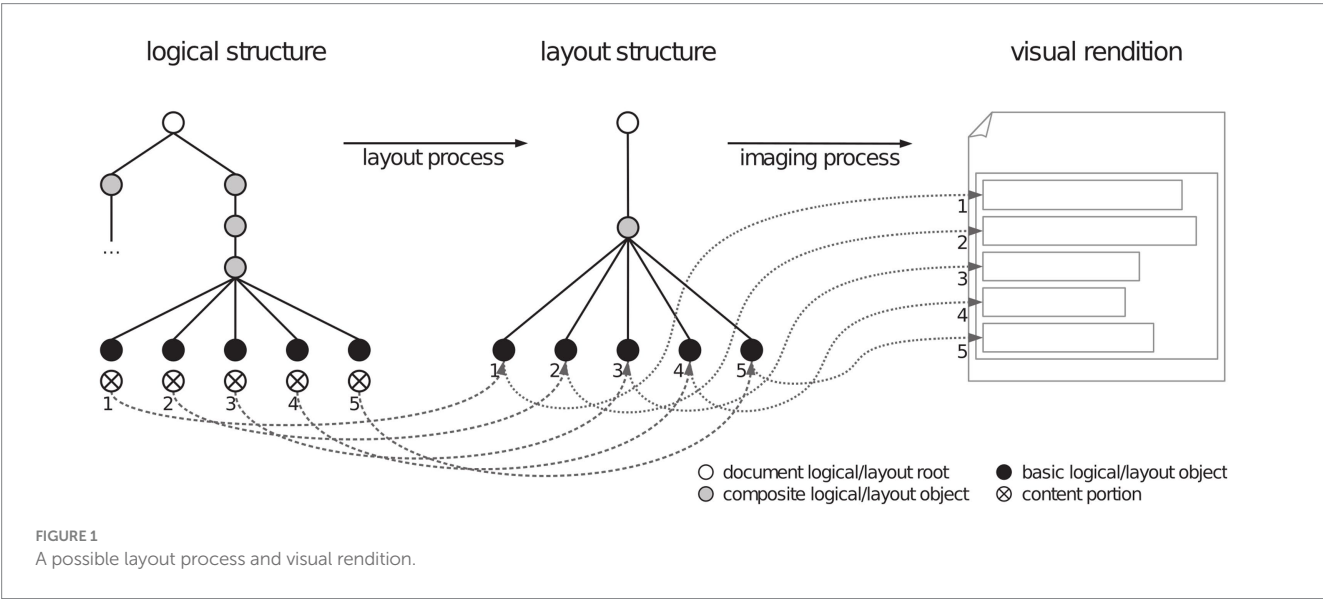
Ιησοῦς  
Χριστός  
Θεῦ  
Υἱός  
Σωτηρ.

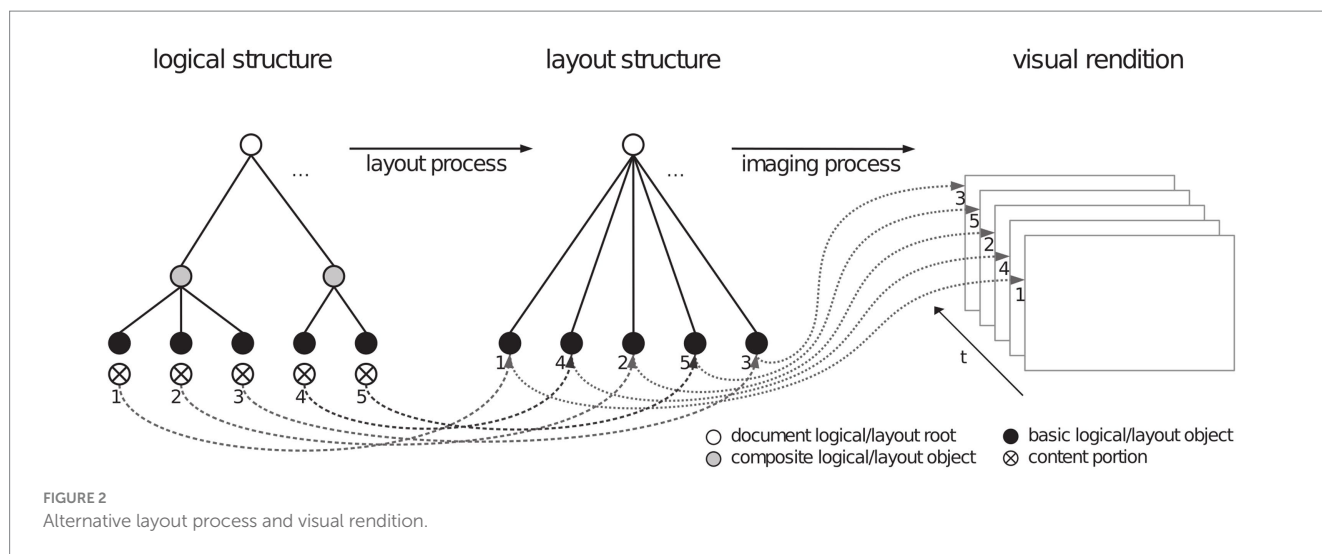
In the first instance, this yields a TEI document (see TEI Consortium, 2023), given in truncated form in the following listing that follows the structure outlined in Figure 1.

```
01 <?xml version="1.0" encoding="UTF-8"?>
02 <TEI xmlns="http://www.tei-c.org/ns/1.0">
03   <teiHeader> ...</teiHeader>
04   <text>
05     <body>
06       <lg type="acrostic">
07         <l>Ιησοῦς</l>
08         <l>Χριστός</l>
09         <l>Θεῦ</l>
10         <l>Υἱός</l>
11         <l>Σωτηρ</l>
12       </lg>
13     </body>
14   </text>
15 </TEI>.
```

The document logical root (<TEI>) here is followed first by two composite logical objects (<teiHeader> and <text>). The <text> element contains further elements: the composite logical objects <body> and <lg>. <lg> is followed by five basic logical objects (<l> in each case). Thus, in macro-navigational terms (i.e., in the sense of identifying parts of the document’s tree structure), five <l> lines below <lg> can be singled out.

A first general possibility for assigning the content portions of the logical structure to the basic layout objects of the layout structure in a layout process can be formulated thus:





**Definition 1.** A (sub)document has a *document-order layout* if the associated set of content portions can be assigned to a set of basic layout objects in such a way that those objects are arranged (spatially and/or temporally) in the logical order.

For electronic documents in the world of XML, the default positioning of layout objects in CSS is sufficient to meet this requirement.<sup>3</sup> Figure 1 shows a document-order layout for the five content portions of the logical objects (lines) of the five-line text in the listing; this can be seen from the dotted arrows that map out the ordering.

For many documents, the requirement of a document-order layout will be too stringent; this is also true of many video documents consisting of several alternating shots, as is familiar from dialogs or car chases. Taking up the textual logical structure from Figure 1, an alternation with five basal logical objects and five associated content portions is illustrated in Figure 2. Here, the content portions numbered 4 and 5 are inserted between the content portions numbered 1 and 2, and 2 and 3, respectively. The logical structure is not preserved. Here, however, there is a link between the logical structure and the layout, satisfying Definition 2, which is weaker than Definition 1.

**Definition 2.** A (sub)document has a *basic-order layout* if the associated set of content portions is assigned to a set of basic layout objects in such a way that the order of all the basic logical objects beneath their respective composite logical objects is preserved.

This means that content portions can—as in the empirically important case of alternation in videos—be rearranged without, for instance, the sequential order specified by the logical structure being lost. In the case of a video segment, the content portions in alternating layouts will typically be shots with their own macro-structures grouping the shots. This will be discussed now.

### 3.2 Structures in video documents

An important question that must be answered in understanding how a presented film works is this: Can a spectator carve up the stream of images rushing past him into meaningful parts? To answer this question, Christian Metz published various studies (Cf. Metz, 1968, 1972, 1974b) from the mid-1960s onwards which dealt with two issues in particular:

- 1 Issues concerning the demarcation of so-called autonomous shots;
- 2 Issues concerning the combination of shots into autonomous 'syntagmatic' forms.

Metz classified some autonomous segments as syntagmatic, providing so-called a-chronological and chronological syntagmas. The so-called *grande syntagmatique* is the classificatory structure that results from the successive dichotomies that organize the syntagmas.<sup>4</sup>

On identifying autonomous segments, Metz writes:

<sup>4</sup> Dudley Andrew offers an early placement of Metz's work within film theory generally (Dudley Andrew, 1976, Chapter 8, pp. 212–241). An early discussion of Metz's approach with respect to the basic semiotic dimensions of language/ langue/parole, form/content/substance, paradigmatic/syntagmatic, etc., can be found in the study by Heath (1973). Good introductions and discussions of the *grande syntagmatique* are given in the study by Stam et al. (1992). Further discussion of subsequent attempts to draw out the paradigmatic and syntagmatic components of the original Metzian scheme can be found in the study by Bateman (2007). In Bateman and Schmidt (2011), chapter 4 "Christian Metz and the *grande syntagmatique* of the image track" is dedicated solely to problems of and critiques raised against the *grande syntagmatique* (Bateman and Schmidt, 2011, pp. 99–128). Important texts of the discussions Metz initiated are translated into English by Buckland (1995); interviews with Metz focusing on his key concepts can be found in the study by Buckland and Fairfax (2017).

<sup>3</sup> This default is found as "normal flow" in, for example, Bos et al. (2011), sec. 9.4.

The analyst of classical film is ... entitled to consider as one (single) autonomous segment any passage of the film which is interrupted neither by a major change in the plot, nor by a punctuation sign, nor by the substitution of one syntagmatic type for another (English quote in [Colin, 1995](#), p. 55).

The problems with this criterion are discussed by Colin in *The Grande Syntagmatique Revisited* ([Colin, 1995](#)). He showed that the notion of a “major change in the plot” is “rather loose” (*ibid.*), that one can be led astray when searching for a punctuation sign in identifying autonomous segments, and that the issues of autonomous segments and of the syntagmatic classification of a film must be treated separately.

This separation can be done by using key concepts of Metz to classify composite logical objects of video documents. A syntagma then classifies such partial trees of the logical structure of a document, which can be rendered in at least one segment.<sup>5</sup> With this aim, the two basic narrative syntagmas, *scene* and *sequence*, are now introduced. Based on these definitions, a concept of alternation is introduced in section 3.3.

In classic alternations such as the dialogs and car chases already mentioned, shots as content portions depict space–time regions and their goings-on. For the five shots assumed in [Figure 2](#), this can happen in very different ways: 5 shots may depict up to 5 different spaces; furthermore, 5 shots may depict a temporal continuity or represent up to 4 temporal gaps. The actual distribution of these conditions results in major differences in the possible structures of a video document concerning the core of their construction, which we will now discuss for any number of several shots.

First, we will conceptualize the minimal situation of a single continuous space–time. If at least two shots depict only a single space and also represent a single temporal continuity, a given number of shots can be assigned to a composite logical object in a document tree and classified as a scene. For this, however, two further conditions must be met, as required by the following Definition 3.<sup>6</sup>

**Definition 3.** A sub-tree of the logical structure of a video document to which at least two shots are assigned as content portions is a **scene** for some set of viewers if:

- 1 the diegetic spaces of all shots assigned to the sub-tree can be conceptualized by all viewers as being connected;
- 2 the diegetic times portrayed in the shots can be conceptualized by all viewers as being connected;
- 3 a layout process exists such that the order of shots created and their diegetic succession can be seen as homomorphic by all

- viewers—i.e., the shots can be displayed in an order that corresponds to the unfolding of events in the diegetic world;
- 4 no further shot meeting conditions (1)–(3) exists.

The last condition expresses an implicit maximality criterion: scenes are maximal because the inclusion of a further shot in a scene is not permitted to result in anything but a scene.

In creating video documents, when representing only one spatial region, unimportant parts of events are often omitted. The temporal continuity of the representation is then deliberately eschewed. This directly results in the following Definition 4 of a sequence (*cf.* [Bateman and Schmidt, 2011](#), p. 210).

**Definition 4.** A sub-tree of the logical structure of a video document to which at least two shots are assigned as content portions is a **sequence** for some set of viewers if:

- 1 the diegetic spaces of all shots assigned to the sub-tree can be conceptualized by all viewers as being connected;
- 2 the diegetic times portrayed in the shots cannot be conceptualized by all viewers as being connected;
- 3 a layout process exists such that the order of shots created and their diegetic succession can be seen as homomorphic by all viewers;
- 4 no further shot meeting conditions (1)–(3) exists.

### 3.3 Alternation

If more than one space is to be depicted in a video segment, the scenes or sequences created may be layouted in alternation. Here, it is important to note that we can only speak of an alternation for a given layout structure. Alternation as a classification only applies to segments of the layout structure, not to the logical structure, because it necessarily involves the commitment to a specific layout.<sup>7</sup> This means that alternation is only weakly dependent on the specific logical structure of the document: the logical structure must, of course, support the creation of an alternation via the layout process but does not itself include that alternation. We use the following Definition 5.<sup>8</sup>

**Definition 5.** A segment in a video document is **n-alternating** with respect to a given layout process and a set of viewers if a partition of the segment exists with *n* partition sets such that:

- 1 the segment consists of shots from scenes or sequences;
- 2 for each pair of partition sets, transitions exist for which a specific symmetric relation holds between some member of the first partition set and some member of the second partition set for all

<sup>5</sup> For the chronological syntagmas of Metz, this was initially developed in articles published by [Schmidt and Strauch \(2002\)](#), [Schmidt \(2004\)](#), [Schmidt \(2008\)](#) and refined in the study by [Bateman and Schmidt \(2011\)](#). An analysis of a longer silent film using these methods can be found in the study by [Bateman and Schmidt \(2011\)](#), pp. 245–286. A toolkit for the analysis of non-syntagmatic autonomous segments with continuous events can be found in the study by [Schmidt and Becher \(2017\)](#).

<sup>6</sup> This definition, refining Metz’s syntagmatic analysis, is found in the study by [Bateman and Schmidt \(2011\)](#), pp. 206 *et seqq.*

<sup>7</sup> Metz introduced alternation as a syntagmatic structure. Problems of this approach are discussed in detail by [Gaudreault and Gauthier \(2018\)](#).

<sup>8</sup> A general definition of alternation is given by [Bateman and Schmidt \(2011\)](#), p. 297.



viewers of the viewer set, and this relation holds for all transitions between the first and second set—this relation then constitutes the coherence for those viewers of the alternating shots it relates;

- 3 for all transition pairs between members of the partition, all viewers of the viewer set conceive the source space–time regions of the members of those pairs to be disjointed;
- 4 in the representation according to the given layout process, there are at least three transitions between the distinct members of each pair of partition sets.

The core of this definition is that (parts of) scenes or sequences are interleaved, and a coherence relation can be stated for the interleaved segment (which, in the case of two scenes/sequences, can label the interleaved segment for a viewer).

The definition does not distinguish whether the alternation is purely a means of representation (a so-called internal alternation based solely on the means of representation) or a diegetically grounded alternation (a so-called external alternation). Thus, no distinction is made between alternations in terms of the paradigmatic difference between internal and external relations: “*internal* distinctions are internal to the text, indicating how the text itself organizes its message ‘rhetorically’; *external* distinctions are in contrast ‘outside’ of the text and are what a text is representing or showing. External relations thus construct relations between the ‘world of events’ depicted in the story; internal relations construct relations in the *telling* of the story” (Bateman and Schmidt, 2011, pp. 177 *et seq.*). The question now is whether this distinction is based solely on assumptions on the part of a viewer or whether it is possible to state criteria for it. This will now be discussed on an empirical basis by analyzing and comparing two very different layouts for the logical structure of the silent film *The Sunbeam*. Both layouts lead to presentations of the video material that are particularly suited to micro-analyses of alternating segments. On this basis, it is possible to mark predetermined breaking points of a viewer’s understanding of alternating layouts created for scenes and sequences and to carry out an analysis of internal and external alternations.

## 4 The Sunbeam

### 4.1 Original and remake

*The Sunbeam* is a motion picture by David W. Griffith, released on 18 March 1912 by Biograph Company, New York.<sup>9</sup> The picture contains a total of 86 shots. In less than 15 min, they tell the story of a little girl in three apartments and the staircase of a building. The story is summarized by Thompson as follows:

“In the opening, a sick mother dies, and her little girl, thinking her mother is asleep, goes out into the hallways of their working-class apartment building. She tries to find someone to play with, but everyone rebuffs her until she manages to charm two lonely people, a bachelor and spinster, who live opposite each other on the floor below the child’s home.” (Thompson, 2011).

<sup>9</sup> The complete picture can be viewed at <http://www.youtube.com/watch?v=bjCyzy5KqZ4>

There is a remake of this motion picture made by Aitor Gametxo.<sup>10</sup> This remake goes beyond the classical filmic montage, i.e., the possibilities provided by the film for ordering sequences of elements in various ways, placing elements in particular orders for particular effects—appropriately labeled “mise-en-chaine” by Gaudreault (1988), p. 119.<sup>11</sup> In this remake, the layout of the shots is arranged in a two-row, three-column grid, and the original serialization in one output stream is replaced by (in principle)  $2 \times 3 = 6$  output streams organized such that the spatial and temporal diegetic events in the apartment building are represented in a largely homomorphic way spatially and, temporally, largely in the diegetic time of the story. Figure 3 is a screenshot showing the dying mother and the girl Sunbeam at the top left; the bottom center, the bachelor mentioned above; and at the bottom right, the spinster from behind.

Both presentations—the original motion picture and the remake—are the very special result of different layouts, each making the same logically structured core of a dataset visible in its own way. Formally, these may also be described by using the progress made in recent decades in the (machine) processing and evaluation of documents; in particular, the separation, now much better understood conceptually, of the logical structure from the layout of a document can be used illustratively.

The logical structure in the original picture and in the remake is dominated by scenes and sequences made very nicely apparent in the remake discussed here, which forms the logical backbone of both layouts. In the layouts of both variants, pivotal points in the diegetic progression can also be identified (see further below). These are predetermined breaking points on the actual reading pathway of a viewer (which may be a machine). Where they occur in alternating use between two spatial regions, they also provide a criterion for distinguishing between internal alternations (which are due to the telling of the story) and external (diegetically grounded) alternations.

Terminological note: Griffith’s original will be referred to as *The Sunbeam*, and the variation by Gametxo will be referred to as *Variation on The Sunbeam* or, for short, as (Gametxo’s) *remake*. In the picture, the nameless spinster is already mentioned, and the nameless bachelor is also already mentioned to become a couple. We will refer to them as *Bachelor* and *Spinster* as proper names. The little girl, as the main protagonist, is called *Sunbeam*.

### 4.2 Phenotypes at the level of the shot

The 86 shots of *The Sunbeam* may be categorized into four phenotypes. These four types are either of a textual nature or genuine cinematic shots identifiably depicting a space–time. Both the title cards (hereinafter T) and the genuine cinematic shots (hereinafter S) will here be numbered 1 through 86 in the order of their appearance in Griffith’s original, the numbers being initially appended to T or S as indices, thus: T<sub>1</sub>, T<sub>2</sub>, S<sub>3</sub>, ..., S<sub>85</sub>, T<sub>86</sub>.<sup>12</sup> There are:

<sup>10</sup> The remake is available at <http://vimeo.com/22696362>

<sup>11</sup> Translated as “putting in sequence” in the study by Gaudreault (2009), p. 91.

<sup>12</sup> This numbering has no theoretical significance. The analogy to the playing-cards metaphor is that a deck of cards is simply numbered consecutively in the order found.

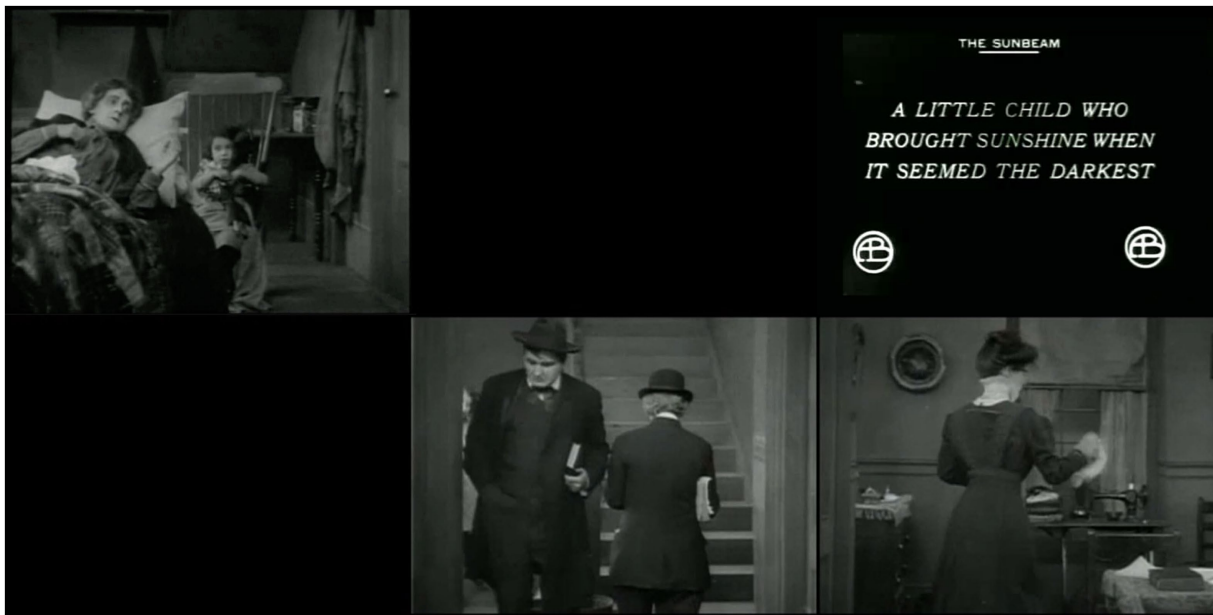


FIGURE 3  
A view of the remake of *The Sunbeam* (from Thompson, 2011).

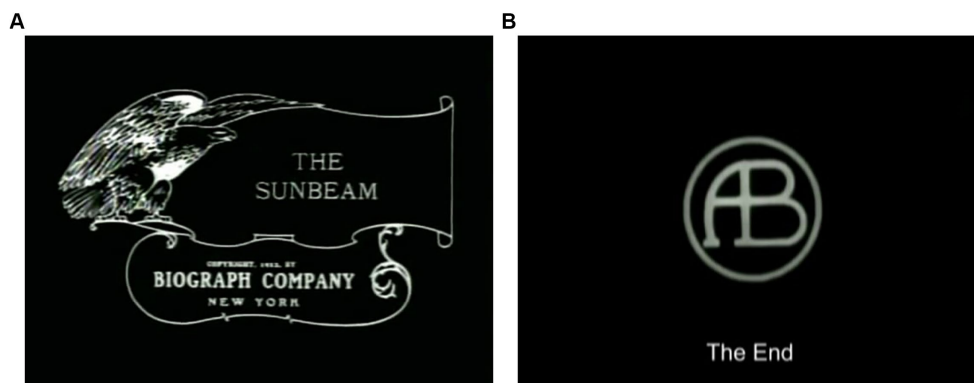


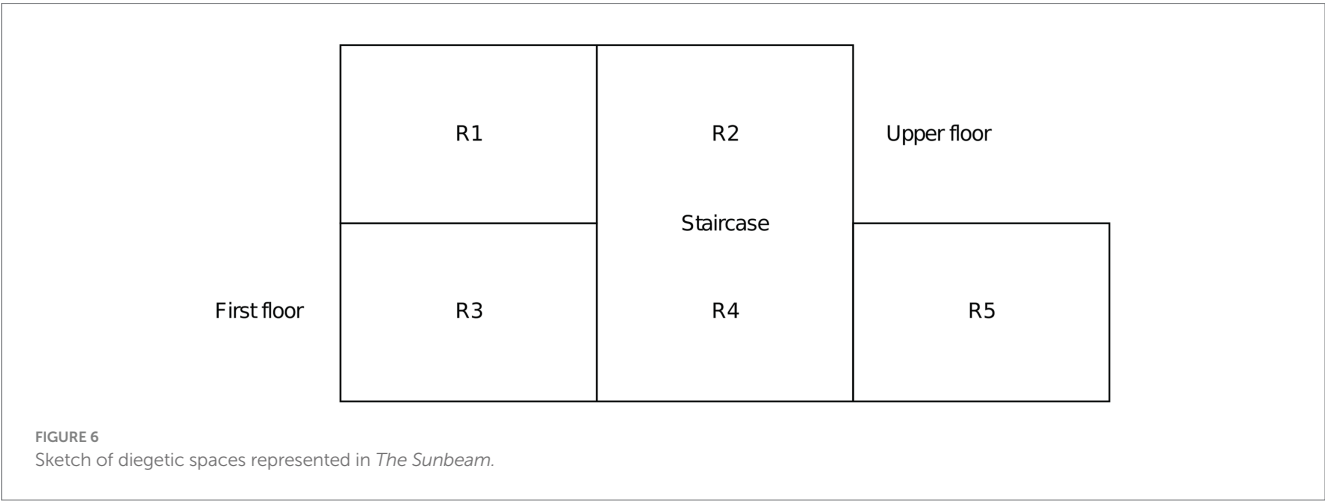
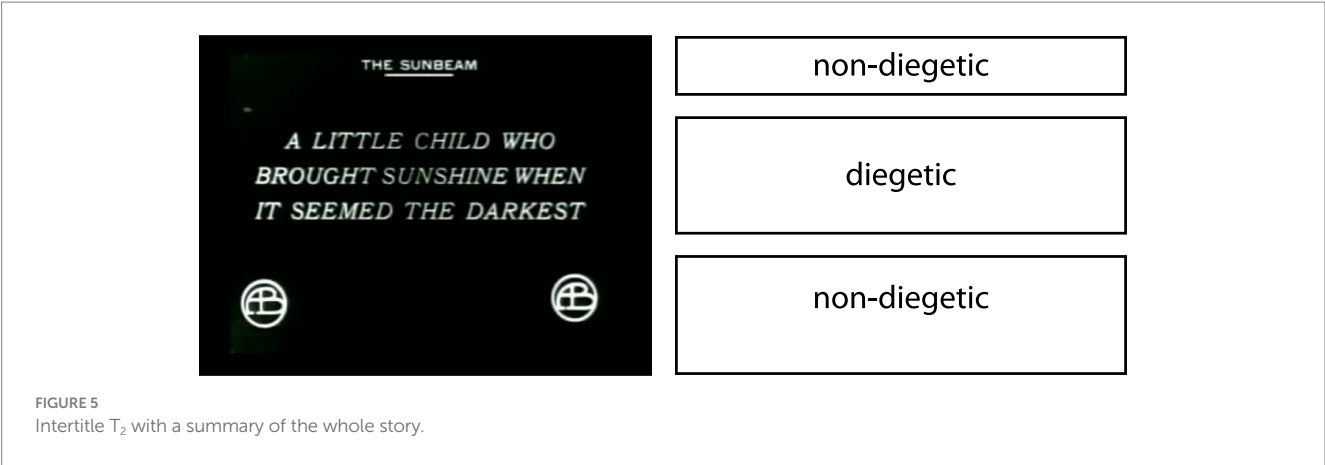
FIGURE 4  
(A,B) Opening title  $T_1$  and closing title  $T_{86}$ .

- 1 An opening title (Type\_1). This type occurs only once in the layout chosen by Griffith, in the first position  $T_1$  of the complete document and.
- 2 A closing title (Type\_2). This type also occurs only once in Griffith's layout, at the end of the complete document in position  $T_{86}$ . Figures 4A,B show these two types.
- 3 A further nine intertitles (Type\_3): Type\_3 is found in the 9 intertitles  $T_2$ ,  $T_4$ ,  $T_{10}$ ,  $T_{36}$ ,  $T_{40}$ ,  $T_{45}$ ,  $T_{53}$ ,  $T_{73}$ , and  $T_{84}$ . The text and the line at the top and the logos at the bottom belong to the non-diegetic content of the document. The rest is diegetically related to each occurrence, showing a particular text in each case. In  $T_2$ , at the very beginning, we find a summary of the story (cf. Figure 5).  $T_2$  is also found at the top right in Figure 3.
- 4 All other shots from  $S_3$  through  $S_{85}$  (75 in all) are not mentioned under points 1–3, the content of which is photographic representations of spatiotemporal events (Type\_4). A key

frame of the first shot of this type in Griffith's original ( $S_3$ ) is found at the top left in Figure 3, where, at the very beginning, the death of the little girl Sunbeam's mother is depicted (see the above summary by Thompson). In the bottom row of Figures 2, 3 further shots of this type are found.

Using shots of Type\_4, the diegetic events are developed in space and time. In what follows, we will, therefore, refer to Type\_4 shots as (*diegetic*) shots. Looking at the spatial regions represented in Type\_4 shots, we find 5 diegetic spaces: three rooms and two parts of the staircase (not separated from each other) of the working-class apartment building, the arrangement of which may be roughly sketched in Figure 6.

To make clear, in what follows, which of these five diegetic spaces, R1 through R5, is shown in the diegetic shots, we will have the numbering of the diegetic shots preceded by the diegetic space



depicted. Accordingly, shots of space R1 will be labeled  $S1_x$ , those of space R2 will be labeled  $S2_x$ , and so on; thus, shot  $S_3$  from the top left of Figure 3 will now be  $S1_3$ .

## 5 Constructing the story

### 5.1 The constitution of space

The space R1, where the little girl (diegetically) lives with her mother, is depicted in the 5 shots  $S1_3$ ,  $S1_{17}$ ,  $S1_{35}$ ,  $S1_{83}$ , and  $S1_{85}$ . Of these, shots  $S1_3$ ,  $S1_{83}$ , and  $S1_{85}$  are part of the cinematic framework of Griffith's original. This framework further includes various title cards, specifically:

- the first title of Type\_1 in  $T_1$ , as shown in Figure 4A above;
- the second title of Type\_3 in  $T_2$ , as shown in Figure 5 above;
- the title of Type\_3 in  $T_{84}$ , as shown in Figure 7B;
- the last title of Type\_2 in  $T_{86}$ , as shown in Figure 4B above.

The diegetic beginning in ( $S1_3$ ,  $S1_{17}$ , and  $S1_{35}$ ) following  $T_2$  is visualized in Figures 8A–C using key frames. Here, a viewer learns early on how the girl Sunbeam is present at her mother's death (she dies as early as in  $S1_3$ ) without registering it and finally leaves the apartment at the end of  $S1_{35}$  so as not to wake her mother. She makes her way into the

core of the building. She then manages “to charm two lonely people, a bachelor and spinster” [see above and Thompson (2011)].

At the end of the picture, we see how Spinster and Bachelor, both charmed by the little girl during the remainder of the picture, together discover the death of the mother and decide to look after the motherless child (“her problem” in  $T_{84}$ ) together (thus solving their problem of loneliness). The ending in ( $S1_{83}$ ,  $T_{84}$ , and  $S1_{85}$ ) is visualized in Figures 7A–C.

The two closing diegetic shots,  $S1_{83}$  and  $S1_{85}$ , though interrupted by an intertitle in Griffith's original, can be played back-to-back without difficulty—as is done by Gametxo in his remake. We will mark such a series of shots, representing an action without interruption, as a single spatiotemporal continuity by underlining them and will refer to such segments as *scenic*. Thus, we have segment ( $S1_{83}$ ,  $S1_{85}$ )—concluding Griffith's original—as a scenic final segment.

When using this underlining convention, we consider the three introductory diegetic shots  $S1_3$ ,  $S1_{17}$ , and  $S1_{35}$  from R1; it is clear that these, too, may be classified as a scene within the above meaning and can be played back-to-back without difficulty. Accordingly, this is what Gametxo does for the scenic segment ( $S1_3$ ,  $S1_{17}$ ,  $S1_{35}$ ) in his remake as well. Here, it becomes apparent from the numbering that the layout of the picture differs significantly from its diegetic progression: The introduction is spread out as far as the 35th shot, which occurs well over one-third into the picture, in only three shots as touchdowns. Thus, it is apparent here that the term scene is a term for the logical structure of a

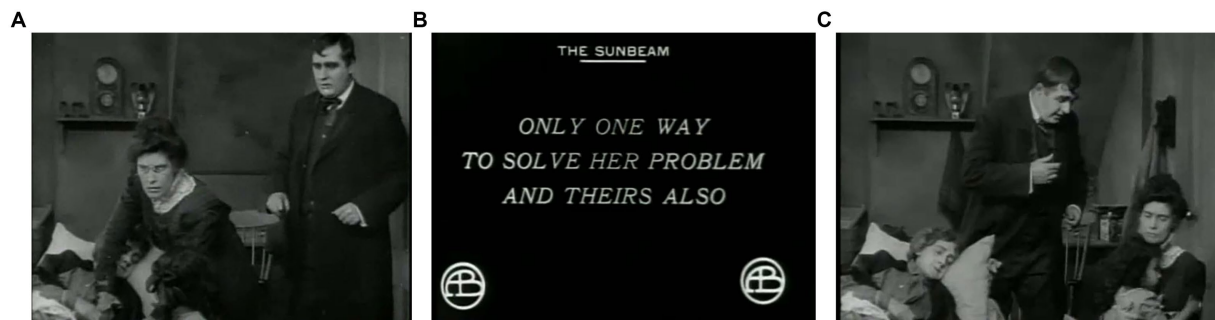


FIGURE 7  
(A–C) Key frames of the final scene ( $S_{183}$ ,  $S_{185}$ ) with intertitle  $T_{84}$ .



FIGURE 8  
(A–C) Key frames of the first scene ( $S_{13}$ ,  $S_{17}$ ,  $S_{135}$ ).

document, which may be assigned a default layout (play back-to-back); however, selecting such a layout is by no means obligatory.

In the original picture, Griffith radically and systematically runs through this latter option, as will now be shown by the following further analysis. To that end, we will discuss the other four spaces, R2 through R5, where the events leading to the happy ending depicted take place, with their corresponding shots. Here, the following quickly becomes apparent: In his original edit, Griffith never depicts the same space in two consecutive shots. In the following lists of shots, there is always at least a distance of 2 between identical space numbers. Thus, the whole picture is—in Griffith's telling of the story—systematically layouted in spatially alternating segments.

The staircase in R2 and R4 is represented in a total of 35 shots—accounting for nearly half of the 75 diegetic shots. While the staircase is represented as two partial diegetic regions, spaces R2 and R4, these will be conceptualized by any normal human viewer as adjoining without separation: in no shot is there a visible separation between them (but nor is there an overlap). The upper space R2 is depicted in the 5-shot segment ( $S_{230}$ ,  $S_{237}$ ,  $S_{237}$ ,  $S_{275}$ , and  $S_{282}$ ). In Figure 9, we show a key frame of shot  $S_{230}$ , in which we see Spinster (still rather grumpily) looking back downstairs. The lower portion of the staircase, R4, is depicted in the 30 shots  $S_{46}$ ,  $S_{48}$ ,  $S_{411}$ ,  $S_{413}$ ,  $S_{415}$ ,  $S_{418}$ ,  $S_{421}$ ,  $S_{424}$ ,  $S_{429}$ ,  $S_{431}$ ,  $S_{433}$ ,  $S_{438}$ ,  $S_{441}$ ,  $S_{443}$ ,  $S_{446}$ ,  $S_{448}$ ,  $S_{450}$ ,  $S_{454}$ ,  $S_{456}$ ,  $S_{458}$ ,  $S_{460}$ ,  $S_{462}$ ,  $S_{464}$ ,  $S_{466}$ ,  $S_{468}$ ,  $S_{470}$ ,  $S_{472}$ ,  $S_{477}$ ,  $S_{479}$ , and  $S_{481}$ . A key frame of the first shot,  $S_{46}$ , can be found in the middle of the bottom row in Figure 3.

Space R3 is the Bachelor's room, depicted in the 24 shots  $S_{37}$ ,  $S_{39}$ ,  $S_{316}$ ,  $S_{320}$ ,  $S_{233}$ ,  $S_{326}$ ,  $S_{328}$ ,  $S_{334}$ ,  $S_{347}$ ,  $S_{349}$ ,  $S_{351}$ ,  $S_{353}$ ,  $S_{355}$ ,  $S_{357}$ ,  $S_{359}$ ,  $S_{361}$ ,  $S_{363}$ ,

$S_{365}$ ,  $S_{367}$ ,  $S_{369}$ ,  $S_{371}$ ,  $S_{374}$ ,  $S_{376}$ ,  $S_{378}$ , and  $S_{380}$ . In Figure 10, we show a key frame of shot  $S_{37}$ , which in Griffith's original is also the first time we are able to look into this apartment.

Space R5 is Spinster's apartment. It is depicted in the 11 shots  $S_{55}$ ,  $S_{512}$ ,  $S_{514}$ ,  $S_{519}$ ,  $S_{522}$ ,  $S_{525}$ ,  $S_{527}$ ,  $S_{539}$ ,  $S_{542}$ ,  $S_{544}$ , and  $S_{552}$ . A key frame of the first shot of R5,  $S_{55}$ , and with Spinster can be found at the bottom right in Figure 3. We can tell from the indices here that this space will have no further role in the latter part of Griffith's original, having been left by the protagonist in the diegetic world in  $S_{552}$ .

Overall, the 86 shots are grouped into 2 + 9 + 75 shots: Alongside the opening (Type\_1) and closing (Type\_2) titles, there are 9 intertitles (Type\_3). The remaining 75 shots can in turn be grouped into 5 groups, depicting 5 spatial regions in 5 + 5 + 24 + 30 + 11 shots (Type\_4), as shown in Figure 11. This very distribution and grouping is the starting point for the remake by Aitor Gametxo.

## 5.2 Continuous spatiotemporal regions: scenes

Of initial importance for the logical structuring of the diegetic shots of spaces R1, R2, R3, R4, and R5 in *The Sunbeam* are those series of shots that can be played back-to-back unbroken in a default layout—as indeed they are in their respective grid window in Gametxo's remake, thus actually depicting diegetic spatiotemporal continuities. These seamlessly depicted series form the scenic segments of the remake, as already discussed in the case



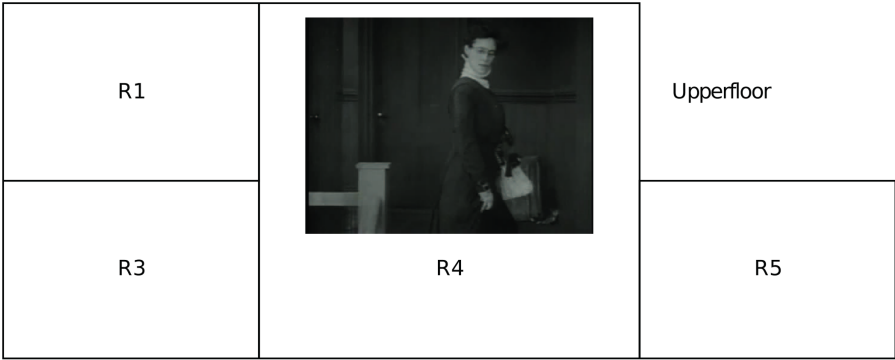


FIGURE 9  
Key frame of shot S<sub>230</sub> with Spinster looking back “downstairs.”



FIGURE 10  
Key frame of shot S<sub>37</sub> with Bachelor.

- R1: (S<sub>13</sub>, S<sub>117</sub>, S<sub>135</sub>), (S<sub>183</sub>, S<sub>185</sub>);
- R2: (S<sub>230</sub>, S<sub>232</sub>, S<sub>237</sub>)<sup>13</sup>;
- R3: (S<sub>23</sub>, S<sub>26</sub>), (S<sub>51</sub>, S<sub>55</sub>, S<sub>57</sub>, S<sub>59</sub>, S<sub>61</sub>, S<sub>63</sub>, S<sub>65</sub>), (S<sub>69</sub>, S<sub>71</sub>, S<sub>74</sub>), (S<sub>76</sub>, S<sub>78</sub>);
- R4: (S<sub>4</sub>, S<sub>8</sub>, S<sub>11</sub>, S<sub>13</sub>, S<sub>15</sub>), (S<sub>18</sub>, S<sub>21</sub>, S<sub>24</sub>), (S<sub>29</sub>, S<sub>31</sub>, S<sub>33</sub>), (S<sub>38</sub>, S<sub>41</sub>), (S<sub>46</sub>, S<sub>48</sub>, S<sub>50</sub>), (S<sub>54</sub>, S<sub>56</sub>, S<sub>58</sub>, S<sub>60</sub>, S<sub>62</sub>, S<sub>64</sub>);
- R5: (S<sub>19</sub>, S<sub>22</sub>, S<sub>25</sub>), (S<sub>42</sub>, S<sub>44</sub>).

Thus, across the whole picture of *The Sunbeam*, there are at least  $2 + 1 + 4 + 6 + 2 = 15$  spatiotemporal regions represented in more than one shot, representing unbroken progressions in the five diegetic spaces.

5.3 Partitioning of diegetic progression: sequences

All scenes in *The Sunbeam*—both in Griffith’s original and in the remake—are embedded in 5 larger sequences representing the diegetic progression in each of the 5 diegetic spaces. In general, a sequence differs from a scene in that the unification of the points in time denoted in the shots is not conceptualized as continuous by a reference viewer. Thus, somewhere between the shots, there is at least one temporal gap. Sequences may contain scenes as temporally unbroken parts. In addition, the series of shots in the logical structure and the diegetic progression must be such that they can be ordered homomorphically by a classifying reference viewer. All this is present here in the sets of shots assigned to the 5 diegetic spaces.

5 shots in R1	5 shots in R2	Intertitles
24 shots in R3	30 shots in R4	11 shots in R5

FIGURE 11  
Distribution of diegetic shots (Type\_4) depicting 5 spatial regions.

of R1. The following scenes rendered as scenic segments in the remake are assumed (structured according to diegetic spaces without underlining):

13 Perhaps only (S<sub>230</sub>, S<sub>232</sub>). In the remake, S<sub>237</sub> is played immediately following S<sub>232</sub>, so that a scenic interpretation is possible: Sunbeam enters the staircase immediately upon Spinster’s leaving it. Whether this is precisely the case is not critical for the purposes of this analysis. The same is true with minor gaps in the representation of the other diegetic spaces.

The whole picture *The Sunbeam* thus contains exactly 5 sequences in which the respective diegetic spaces are represented and which, in the remake, are played in their respective grid cells (the embedded scenes leading to scenic segments in the remake are underlined):

- The R1 sequence (S1<sub>3</sub>, S1<sub>17</sub>, S1<sub>35</sub>, S1<sub>83</sub>, S1<sub>85</sub>): there are 2 continuous temporal regions and thus exactly one temporal gap;
- The R2 sequence (S2<sub>30</sub>, S2<sub>32</sub>, S2<sub>37</sub>, S2<sub>75</sub>, S2<sub>82</sub>) or (S2<sub>30</sub>, S2<sub>32</sub>, S2<sub>37</sub>, S2<sub>75</sub>, S2<sub>82</sub>): there are 3 (or, depending on viewers' preferences, See "Footnote 13", 4) temporal regions, one scene and 2 (or 3) temporal gaps;
- The R3 sequence (S3<sub>7</sub>, S3<sub>9</sub>, S3<sub>16</sub>, S3<sub>20</sub>, S3<sub>23</sub>, S3<sub>26</sub>, S3<sub>28</sub>, S3<sub>34</sub>, S3<sub>47</sub>, S3<sub>49</sub>, S3<sub>51</sub>, S3<sub>55</sub>, S3<sub>57</sub>, S3<sub>59</sub>, S3<sub>61</sub>, S3<sub>63</sub>, S3<sub>65</sub>, S3<sub>67</sub>, S3<sub>69</sub>, S3<sub>71</sub>, S3<sub>74</sub>, S3<sub>76</sub>, S3<sub>78</sub>, S3<sub>80</sub>): there are 14 temporal regions (with 4 scenes) and thus 13 gaps.
- The R4 sequence (S4<sub>6</sub>, S4<sub>8</sub>, S4<sub>11</sub>, S4<sub>13</sub>, S4<sub>15</sub>, S4<sub>18</sub>, S4<sub>21</sub>, S4<sub>24</sub>, S4<sub>29</sub>, S4<sub>31</sub>, S4<sub>33</sub>, S4<sub>38</sub>, S4<sub>41</sub>, S4<sub>43</sub>, S4<sub>46</sub>, S4<sub>48</sub>, S4<sub>50</sub>, S4<sub>54</sub>, S4<sub>56</sub>, S4<sub>58</sub>, S4<sub>60</sub>, S4<sub>62</sub>, S4<sub>64</sub>, S4<sub>66</sub>, S4<sub>68</sub>, S4<sub>70</sub>, S4<sub>72</sub>, S4<sub>77</sub>, S4<sub>79</sub>, S4<sub>81</sub>): there are 14 temporal regions (with 6 scenes) and thus 13 gaps;
- The R5 sequence (S5<sub>5</sub>, S5<sub>12</sub>, S5<sub>14</sub>, S5<sub>19</sub>, S5<sub>22</sub>, S5<sub>25</sub>, S5<sub>27</sub>, S5<sub>39</sub>, S5<sub>42</sub>, S5<sub>44</sub>, S5<sub>52</sub>): for R5, there are 8 temporal regions (with 2 scenes) and thus 7 temporal gaps.

These sequences structure the diegetic shots, jointly creating the maximal chronological partial document (cf. Bateman and Schmidt, 2011, p. 205) of the document *The Sunbeam*. In contrast to natural languages with local sentence structures, sequences can extend over the whole video document and can, in principle, continue to alternate, interleaved in their alternation up to the end. In rendered video documents, the end of a sequence is only reached when its space is depicted for the last time: in the case of the remake of *The Sunbeam*, the respective grid cell in the "dollhouse" (cf. above and Thompson, 2011) is vacated after such a last shot.

## 5.4 The basal structure of *The Sunbeam*

The whole structure of *The Sunbeam*, then, is based on 75 shots of Type<sub>4</sub> contained in 5 sequences. These contain a total of  $2 + 3(4) + 14 + 14 + 8 = 41$  (or 42) spatiotemporal regions represented continuously, of which  $2 + 1 + 4 + 6 + 2 = 15$  are represented scenically in more than one shot in the remake.

No two diegetic shots from one diegetic space follow each other immediately in Griffith's original layout—not even where they can scenically represent one process, meaning they can be played back-to-back without difficulty in Gametxo's remake. The progression of the original picture is thus subject to constant changes. This also applies to the 15 scenes underlined above. This shows how far Griffith had departed from stage conceptions as early as 1912: he even rips apart possible scenic segments as a matter of principle. In particular, the opening scene in the mother's death chamber is, in Griffith's original, drawn out far into the motion picture. This scene, and all others, are reassembled in their default layout in Gametxo's remake, with the shots in their spaces played consecutively. To put it simply, Gametxo re-stages Griffith's dramatic composition in his remake.

The common conceptual starting point of the structures of scene and sequence (cf. Definitions 3 and 4) is their reference to a measurable

spatial unit in the shots. Where an action diegetically transcends spatial regions in two shots adjacent to the layout, there is a spatiotemporal transition. For a viewer, special singular spatiotemporal bridges then function as predetermined cognitive breaking points for understanding the diegetic progression, as will now be shown.

## 5.5 The backbone of the story: sunbeam's itinerary

To reconstruct the story of the picture, we use the itinerary of the heroine Sunbeam in the original picture *The Sunbeam* and in the remake. This itinerary forms the backbone of the whole picture and is restrictive in the following way: Any part of a shot that includes a depiction of this itinerary cannot diegetically overlap with a part of another shot that includes a depiction of this itinerary. In the remake, this becomes apparent because we can follow Sunbeam without difficulty, as she only ever appears in at most one grid cell. Sunbeam is represented in the following segment, an explanation of which (including the bold emphasis) follows:

S1<sub>3</sub>, S1<sub>17</sub>, S1<sub>35</sub>, //Sunbeam is present at the death of her mother without noticing it.

S2<sub>37</sub>, //Sunbeam enters the upper part of the staircase.

S4<sub>38</sub>, S4<sub>41</sub>, //Sunbeam is in the lower part of the staircase.

S5<sub>42</sub>, S5<sub>44</sub>, //Sunbeam charms Spinster in her apartment.

S4<sub>46</sub>, S4<sub>48</sub>, S4<sub>50</sub>, //Sunbeam is in the lower part of the staircase.

S3<sub>51</sub>, S3<sub>55</sub>, S3<sub>57</sub>, S3<sub>59</sub>, S3<sub>61</sub>, S3<sub>63</sub>, S3<sub>65</sub>, S3<sub>67</sub>, S3<sub>69</sub>, S3<sub>71</sub>, S3<sub>74</sub>, S3<sub>76</sub>, S3<sub>78</sub>, S3<sub>80</sub>

//Sunbeam charms Bachelor, first on his own, then with Spinster present, in his apartment (note: three scenes).

S4<sub>81</sub>, //Sunbeam, carried by Bachelor, is in the lower part of the staircase.

S2<sub>82</sub>, //Sunbeam, carried by Bachelor, is in the upper part of the staircase.

S1<sub>83</sub>, S1<sub>85</sub>, //Sunbeam, along with Bachelor and Spinster, is in her dead mother's apartment. Her death is noticed by Bachelor and Spinster. They decide to look after Sunbeam together.

Sunbeam is thus only seen in 29 of 75 diegetic shots. However, this visible time of Sunbeam covers almost the entire diegetic time of the picture: There are very few time intervals in Gametxo's remake where Sunbeam is not seen at all.<sup>14</sup>

<sup>14</sup> This happens in the representation of shot S<sub>66</sub> (apart from a minor and negligible initial overlap with shot S<sub>65</sub>, where Sunbeam is seen), in the representation of all of S<sub>68</sub>, of a middle portion of S<sub>79</sub>, and the beginning of S<sub>83</sub>.

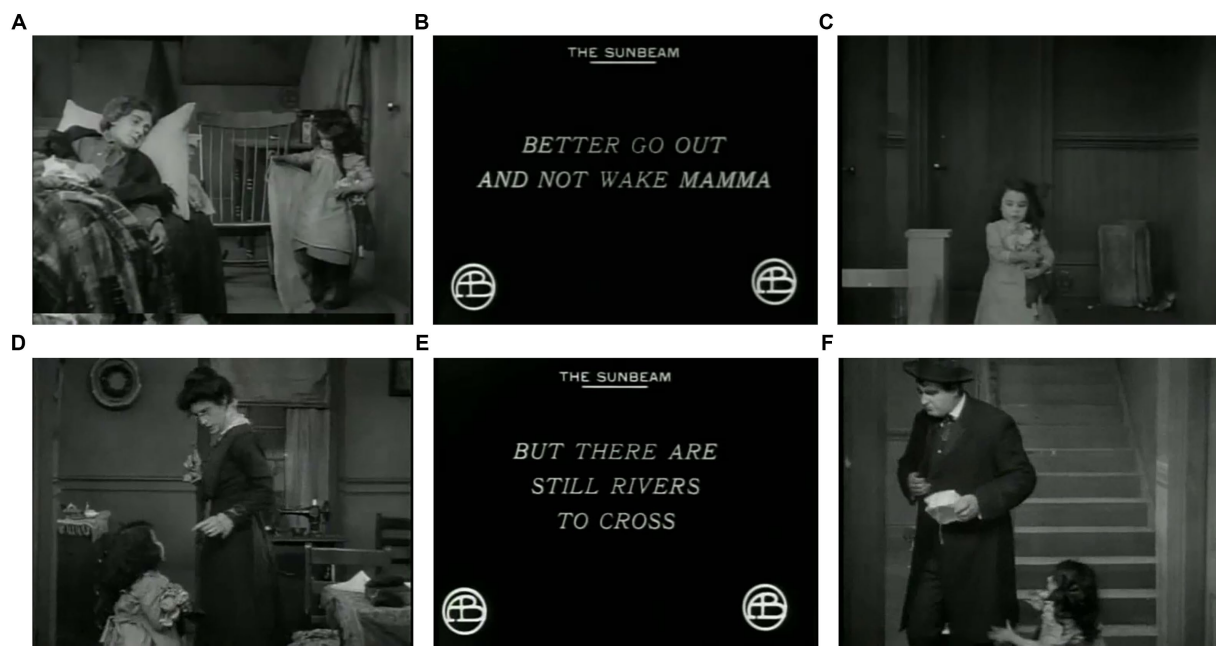


FIGURE 12  
(A–F) Segments (S1<sub>35</sub>, T<sub>36</sub>, S2<sub>37</sub>) and (S5<sub>44</sub>, T<sub>45</sub>, S4<sub>46</sub>) with intertitles T<sub>36</sub> and T<sub>45</sub>.

The entire Sunbeam itinerary contains 8 scenes (again marked by underlining in the above list), which in turn—as is apparent from the above list—cover a large portion of the itinerary. These scenes have no unusual features; in the remake, they are put in their default play back-to-back layout. Anyone who understands these scenic segments will understand a large part of the itinerary.

To understand the full itinerary, it is necessary to master the critical starting and end points of these scenes, and the shots are not part of any scenes. These points are decisive moments for the content of the Sunbeam itinerary and thus of the whole picture. These are the points emphasized in **bold** above. Where two critical points occur in succession, two types are distinguishable:

- There is a spatial transition between two adjacent critical shots; or,
- Two adjacent critical shots are part of the same sequence; thus representing one and the same diegetic space.

For the first case in this list, the following spatial transitions between adjacent shots are present in *The Sunbeam* within the Sunbeam itinerary: S1<sub>35</sub> (T<sub>36</sub>) S2<sub>37</sub>; S2<sub>37</sub>, S4<sub>38</sub>; S4<sub>41</sub>, S5<sub>42</sub>; S5<sub>44</sub> (T<sub>45</sub>) S4<sub>46</sub>; S4<sub>50</sub>, S3<sub>51</sub>; S3<sub>80</sub>, S4<sub>81</sub>; S4<sub>81</sub>, S2<sub>82</sub>; S2<sub>82</sub>, S1<sub>83</sub>. Among these, the first and fourth spatial transitions also include the intertitles T<sub>36</sub> and T<sub>45</sub>, as shown in [Figures 12B,E](#).

For the second case, there are the following critical points emphasized in **bold**: S3<sub>65</sub>, S3<sub>67</sub>, S3<sub>69</sub>, S3<sub>74</sub>, S3<sub>76</sub>, S3<sub>78</sub>, S3<sub>80</sub>. Here, as the

indices and their distances show, a viewer must create an understanding of shots that are not part of Sunbeam's itinerary. This may also involve the beginning of an alternation or an insertion [within the meaning of “broad syntagmatic types” according to [Bateman and Schmidt \(2011\)](#), pp. 171 *et seqq.*].

## 6 Progressive spatial transitions and alternations

### 6.1 Progressive spatial transitions and progression bridges

The comparison of the two layouts of the original version by Griffith and the remake by Gametxo now allows for micro-analyses of the relation between a viewer and the layouted document, leading to a differentiation between internal and external alternations.

A transition from a given spatial region to a different spatial region diegetically represented later can occur from several spatial regions (including diegetically simultaneous ones). We will call these transitions *progressive spatial transitions*. Among these, we will (following the language of graph theory) mark as *progression bridges* those progressive spatial transitions for which diegetic progression can only occur through a transition that is unambiguously defined within the whole document and which a viewer must cross to make any temporal diegetic progress in absorbing the content of the document for a given layout. To put it simply, this is a bridge a viewer must cross to take the next step in understanding the whole document. Conceptually, this requires a preliminary boundary point in diegetic time from which one can only progress through an unambiguously defined transition of diegetic place. These bridges are essential hinge points of a motion picture and predetermined cognitive breaking

negligible in the overall tally. These “Sunbeam-free” phases must be added to the diegetic time of the “Sunbeam” phases to determine the approximate diegetic time of the story.

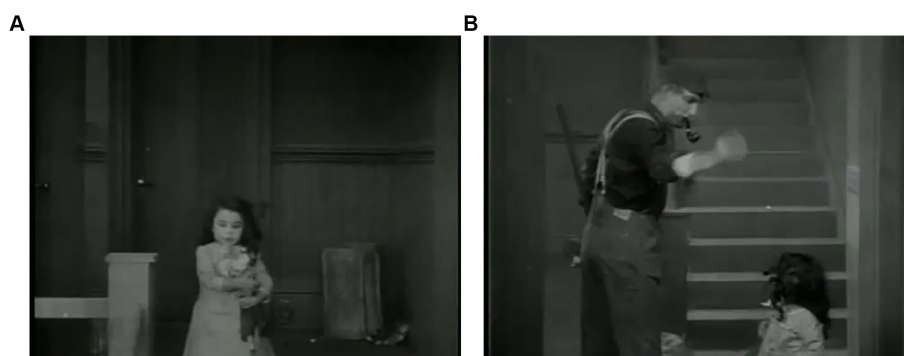


FIGURE 13  
(A,B) The first progression bridge in (S2<sub>37</sub>, S4<sub>38</sub>).

points for understanding the diegetic progression, which we will now show in another round of analysis from the diegetic beginning to the diegetic ending of the picture.

## 6.2 Singular spatial transitions

The first two progressive spatial transitions in the above list in (S1<sub>35</sub>, T<sub>36</sub>, S2<sub>37</sub>) and in (S2<sub>37</sub>, S4<sub>38</sub>) provide a set of examples for distinguishing between a simple progressive spatial transition and a progression bridge: In the first transition, visualized in Figures 12A–C, there is no bridge; in the second transition, there is a bridge. By way of justifying this, the progression from the diegetic beginning in S1<sub>3</sub> up to and including S4<sub>38</sub> will now be sketched in the approximate temporal order of the shots.

The first scene (S1<sub>3</sub>, S1<sub>17</sub>, S1<sub>35</sub>), visualized in Figures 8A–C, covers the entire exposition of the picture in temporal diegetic terms so that more than one-third of the shots in the picture occur during the exposition. The second scene (S2<sub>30</sub>, S2<sub>32</sub>, S2<sub>37</sub>), as the indices show, grows out of this portion. In segment (S1<sub>35</sub>, T<sub>36</sub>, S2<sub>37</sub>), Sunbeam leaves the space R1—with the intertitle “BETTER GO OUT AND NOT WAKE MAMMA” in T<sub>36</sub> (cf. Figure 12B)—and, in S2<sub>37</sub>, enters the upper part R2 of the staircase. In the layout of *The Sunbeam*, this is a progressive spatial transition but not a progression bridge since it is apparently also possible to diegetically progress from S2<sub>32</sub> to S2<sub>37</sub>.

In contrast, there is a bridge in the next transition, from S2<sub>37</sub> to S4<sub>38</sub>, as shown in Figures 13A,B. At the end of S2<sub>37</sub>, the temporal diegetic progression reaches a point where it is only possible to “move on” by a spatial transition to R4, the lower portion of the staircase. The segment (S2<sub>37</sub>, S4<sub>38</sub>) marks the **first progression bridge** (in the whole picture). Sunbeam here transitions to the lower part of the building, linking the exposition to the rest of the diegetic events. This transition must absolutely be understood by a machine or human viewer. Otherwise, the story will disintegrate into 2 components, an “upper” component with Sunbeam and the death of her mother, and a “lower” component with Sunbeam’s attempts at social contact.

This first progression bridge is followed by a **second** progression bridge in (S4<sub>41</sub>, S5<sub>42</sub>) and a **third** in (S5<sub>44</sub>, S4<sub>46</sub>)—in the latter case with an intertitle T<sub>45</sub>, as visualized in Figure 12D–F.

Following her scenically represented stay in the hallway R4 (in Griffith’s original, scene (S4<sub>46</sub>, S4<sub>48</sub>, S4<sub>50</sub>) is interspersed with shots S3<sub>47</sub>

and S3<sub>49</sub> from the R3 sequence), Sunbeam, in a progressive spatial transition in (S4<sub>50</sub>, S3<sub>51</sub>), enters the Bachelor’s apartment in S3<sub>51</sub>. This spatial transition (S4<sub>50</sub>, S3<sub>51</sub>) is not a bridge, as S4<sub>50</sub> seamlessly follows S4<sub>48</sub> in parallel to S3<sub>49</sub> so a diegetically progressive reading path is possible via both the segments (S4<sub>48</sub>, S4<sub>50</sub>, S3<sub>51</sub>) and (S4<sub>48</sub>, S3<sub>49</sub>, S3<sub>51</sub>).

## 6.3 Series of spatial transitions

Having reached space R3, Sunbeam, first on her own with Bachelor and then joined by Spinster, initiates the happy ending. The core structures here are the scenes (S3<sub>51</sub>, S3<sub>55</sub>, S3<sub>57</sub>, S3<sub>59</sub>, S3<sub>61</sub>, S3<sub>63</sub>, S3<sub>65</sub>), (S3<sub>69</sub>, S3<sub>71</sub>, S3<sub>74</sub>) and (S3<sub>76</sub>, S3<sub>78</sub>). With these scenes, Sunbeam’s story in this picture is almost complete. The transition to the actual happy ending from S3<sub>78</sub> only remains.

In the first R3 scene, in Griffith’s original layout, S3<sub>51</sub> is followed by the segment (S3<sub>55</sub>, S3<sub>57</sub>, S3<sub>59</sub>, S3<sub>61</sub>, S3<sub>63</sub>, S3<sub>65</sub>), edited to alternate with the R4 segment (S4<sub>54</sub>, S4<sub>56</sub>, S4<sub>58</sub>, S4<sub>60</sub>, S4<sub>62</sub>, S4<sub>64</sub>), as shown in the tabular sketch of the temporal diegetic relations in Figure 14 (progressing vertically for clarity, with R3 on the left and R4 on the right).

In the given layout, there are no progression bridges here, only progressive spatial transitions.<sup>15</sup> Each shot in this alternation can be reached in at least two ways in the spatiotemporal progression.

From S4<sub>60</sub>, a children’s prank is introduced as a subplot, in which several children affix the sign “SCARLET FEVER,” already shown in S4<sub>43</sub>, to the door of Bachelor’s apartment and then fetch the police—to alert them to Bachelor as a possible epidemic focus and so to annoy him (Bachelor has no friends in the building, as we have already been told by the title card “EVERYBODY HAS FRIENDS BUT HIM” in T<sub>10</sub> in Griffith’s original).

In Griffith’s alternating edit, the layout depicts the temporal diegesis homomorphically according to its progression; however, this depiction is not defined unambiguously. With the same diegesis, various transpositions can be made to the alternating layout while retaining the playback of the diegetic progression, such that the two largely parallel (in temporal diegetic terms) spatiotemporal regions are

<sup>15</sup> Progressive spatial transitions include those where two diegetic times are the same, as in S3<sub>54</sub> and S4<sub>55</sub>, for example.



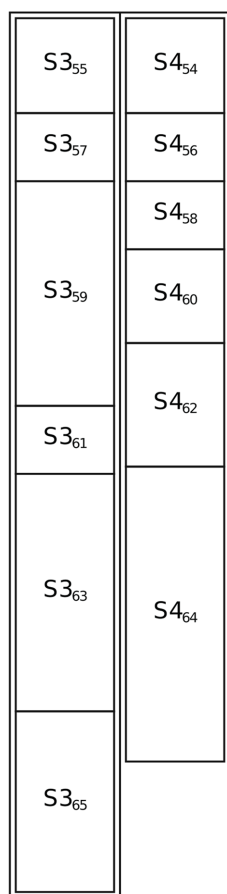


FIGURE 14  
Progressive spatial transitions in (S3<sub>55</sub>, ..., S3<sub>65</sub>).

represented each in the right order. Thus, in the layout, S3<sub>57</sub> could easily be swapped for S4<sub>56</sub>. Thus, there are degrees of freedom that still meet the boundary condition of some diegetic progression. The underlying structure can be alternated such that the telling of the story can represent the diegetic progression in multiple ways.

In contrast, the segment (S3<sub>65</sub>, S4<sub>66</sub>, S3<sub>67</sub>, S4<sub>68</sub>, S3<sub>69</sub>) immediately following diegetically predetermines the alternation. Here, for the first time in this picture, we find a classic alternating layout of the form change of place as time goes on. Whereas, up to now, a viewer was able to go through the picture in scenic segments with occasional spatial transitions, beginning with S3<sub>65</sub>, this is no longer possible: in (S3<sub>65</sub>, S4<sub>66</sub>, S3<sub>67</sub>, S4<sub>68</sub>, S3<sub>69</sub>), for the first time, a viewer definitely cannot progress scenically through the picture but is diegetically compelled to jump back and forth four times between R3 and R4, as shown in Figures 15A–E. What is crucial here: Each individual jump is a progression bridge—we here have the **fourth through seventh progression bridges** of the picture in succession, externally determining an alternating layout. This is—singularly in *The Sunbeam*—an **externally determined alternation**.

From the R3 part of the itinerary, there now remains the final R3 scene (S3<sub>76</sub>, S3<sub>78</sub>), which is linked in alternation to the preceding R3 scene (S3<sub>69</sub>, S3<sub>71</sub>, S3<sub>74</sub>) via S2<sub>75</sub> and to the rest of the picture via S4<sub>77</sub>, S4<sub>79</sub>, and S4<sub>81</sub>. This chosen alternation, however, is not defined by external diegetic conditions. If the underlying structure can be considered suitable for alternation at all (this, after all, requires the

specification of a coherence relation on the part of a viewer), it is only suitable for homomorphic alternation, but the diegesis does not define a default layout for the progression represented.

Sunbeam's whole itinerary ends with her being carried up the stairs and so (by means of the **eighth and ninth progression bridges** in her itinerary from S4<sub>81</sub> to S2<sub>82</sub> and from S2<sub>82</sub> to S1<sub>83</sub>) returning to her dead mother's apartment R1 for the concluding scene, the happy ending (S1<sub>83</sub>, S1<sub>85</sub>).<sup>16</sup> This ends Sunbeam's itinerary through the picture. The picture ends with the closing title card T<sub>86</sub>.

## 6.4 Internal and external alternations

In Gametxo's remake, progression bridges are exactly those points where the forming of connections in understanding the content must necessarily transcend grid cells, moving from one cell to another, in order for the understanding to reach the end of the document. In the given document, none of these points is sensational in content: they are all situations of movement and/or (partial) itineraries. For that reason, they are inconspicuous in Griffith's original. Only in comparison with the remake by Gametxo can they be identified as predetermined breaking points in understanding the diegetic progression of the picture.

By introducing progression bridges, it becomes possible to mark **external alternations**. Where an alternating layout consists solely of such bridges, the underlying bridge structure, through the assumed diegesis, defines an alternating default layout—in the same way as the seamless playback of the shots is the default layout for a scene. If, however, instead of a progression bridge, there is only a progressive spatial jump, no unambiguously defined default layout for the representation of the diegetic progression is predetermined. To the extent to which a structure suitable for alternation is desired, the creator then has liberties in ordering the layout, which can be used in designing **internal alternations** without violating a viewer's temporal diegetic intuitions.

This result takes up an old discussion in Metz himself: Metz, at one point, terms examples of external alternation pseudo alternation to differentiate them from alternation proper as a discourse strategy (Metz, 1974a, p. 164n). The result here is that if an alternation is based on progression bridges, it is an externally based alternation for which a default layout applies. Otherwise, the organization of alternations can be based on reasons internal to the discourse.

## 7 Discussion

The silent film *The Sunbeam* by Griffith and its remake by Gametxo seem to be special in their film-theoretical context, but they

<sup>16</sup> The fact that only bridges are present here and no parallel plot is represented suggests the interpretation that the happy ending should now occur rapidly, without beating about the diegetic bush, as would be suggested by a different spatial transition. Overall, the picture *The Sunbeam* contains 10 bridges: 9 in Sunbeam's itinerary and one more in (S3<sub>78</sub>, S4<sub>79</sub>), leading us out of Sunbeam's itinerary.



FIGURE 15  
(A–E) External alternation in *The Sunbeam* in (S3<sub>65</sub>, S4<sub>66</sub>, S3<sub>67</sub>, S4<sub>68</sub>, S3<sub>69</sub>).

fit seamlessly into today's view of document processing with multimodal content, as described in the introduction.

It is striking that the whole of Griffith's original document, which brims with alternations, actually contains only one externally determined alternation suggested by the diegesis—in the out-of-line segment (S3<sub>65</sub>, S4<sub>66</sub>, S3<sub>67</sub>, S4<sub>68</sub>, S3<sub>69</sub>) with the four progression bridges. This is exactly what gives Griffith the liberty to edit in an extremely alternating fashion and also what makes the external alternation conspicuous. This liberty is made possible by the generally weak conditions in which the logical structure of a video document according to Definition 2 specifies for the layout and thus for the representation of a document. This can also be calculated, as will now be shown for the 75 diegetic shots of *The Sunbeam*.

*A priori*, a document with  $n$  content portions has  $n!$  (i.e.,  $n \times (n-1) \times \dots \times 2 \times 1$ ) possible arrangements of these content portions at  $n$  places in the layout, if no other specifications are made. For content portions obtained from multimodal corpora, further restrictions may apply, resulting from the rules for mapping the logical structure into a layout, as specified, for example, in Definitions 1 and 2.

For a structured text with 75 sentences as 75 content portions, if one is forced to maintain the logical structure in the layout according to Definition 1, there would only be 1 possible solution for this text. For many texts, such a requirement makes sense to preserve the sentence order; for video documents, however, more freedom is often allowed, which quickly leads to a wide range of possibilities for the *mise-en-chaine*.

In *The Sunbeam*, there is a document that does not contain 75 arbitrary diegetic shots but only 5 sequences. The definition of sequences used here stipulates that the associated shots can be arranged in such a way that the order of shots created and their diegetic succession can be seen as homomorphic by all viewers. This, in turn, means that with a layout according to Definition 2, the shots

can be displayed in an order corresponding to the unfolding of events in the respective diegetic space.

The distribution of the number of shots in *The Sunbeam* over the 5 diegetic spaces, R1 to R5, is as follows: There are 5 shots for R1, 5 shots for R2, 24 for R3, 30 for R4, and 11 for R5. If the respective order of these shots is not changed in the layout of the overall document according to Definition 2, the number of the 75! possible arrangements is reduced by  $(5! \ 5! \ 24! \ 30! \ 11!)$ , a reduction of almost  $10^{68}$  possibilities. This leaves a maximum of  $75! / (5! \ 5! \ 24! \ 30! \ 11!)$  different solutions for the diegetic shots only. Despite the significant reduction due to the denominator, this still results in a 42-digit number:

$$2.62257410581244368515476894205849109824 \times 10^{41}.$$

As a result, Griffith had a great deal of freedom for his montage of *The Sunbeam* in an order that satisfied his wishes. However, if the shots are not arranged directly one after the other for each diegetic space (for example, first the R1 shots, then the R2 shots, etc.),<sup>17</sup> shots from different diegetic spaces must necessarily alternate when brought into a chain. This can be done with internal (discursively motivated) or external (diegetically motivated) alternations. It is, therefore, necessary for both production and analysis to look for justified restrictions of possible alternations. A significant restriction was specified here in the identification of external alternations, which determines the layout by default for a number of shots.

<sup>17</sup> The video "Deconstructing Griffith - A Girl and Her Trust (1912) prior to editing" by Jim Middleton even shows the shots of the individual camera positions for the silent film "The Girl and Her Trust," starting with the interior shots and followed by the exterior shots. See <https://www.youtube.com/watch?v=BzGITh-Olg> (16.03.2024).

Gametxo has subjected Griffith's original to detailed analysis at the level of shots. This analysis led to his remake as an independent work. With the current state of document processing, such a variation with its tabular layout can be generated from the same logical structure as Griffith's version. This can be realized by two stylesheet specifications, one for alternating presentation of the 5 sequences and the other for presentation in the 5 table cells which are assigned to the diegetic spaces R1 to R5.<sup>18</sup> The original by Griffith and the remake are to be seen today as different layouts for a screen output of a common logical structure, as introduced in the first three sections of this study.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

<sup>18</sup> This can even be done on the user side alone. In general, a layout is determined by the intentions of the creator (leading to "author styles" in the terminology of CSS), existing and chosen output options (leading to "user agent styles" in the terminology of CSS) and a viewer's needs (leading to "user styles" in the terminology CSS; Cf. Bos, 2016, sec. 6.4).

## References

- Bateman, J. (2007). Towards a grande paradigmatic of film: Christian Metz reloaded. *Semiotica* 2007, 13–64. doi: 10.1515/SEM.2007.070
- Bateman, J., and Schmidt, K.-H. (2011). *Multimodal film analysis: How films mean*. New York, London: Routledge.
- Bos, B. (2016). Cascading style sheets level 2 revision 2 (CSS 2.2) specification. W3C working draft, W3C, April 2016. Available at: <https://www.w3.org/TR/CSS22/> [Accessed March 23, 2024].
- Bos, B., Çelik, T., Hickson, I., and Lie, H.W. (2011) Cascading style sheets level 2 revision 1 (CSS 2.1) specification. W3C recommendation, W3C, June 2011. Available at: <https://www.w3.org/TR/2011/REC-CSS2-20110607/> [Accessed November 2, 2023].
- Buckland, W. (1995). *The film spectator: From sign to mind*. Amsterdam: Amsterdam University Press.
- Buckland, W., and Fairfax, D. (2017). *Conversations with Christian Metz: Selected interviews on film theory (1970–1991)*. Amsterdam: Amsterdam University Press.
- Colin, M. (1995). "The grande syntagmatique revisited" in *The film spectator: From sign to mind*. ed. W. Buckland (Amsterdam: Amsterdam University Press), 45–85.
- Dudley Andrew, J. (1976). *The major film theories: An introduction*. Oxford: Oxford University Press.
- Gametxo, A. (2011). Variation on "The Sunbeam". Available at: <https://vimeo.com/22696362> [Accessed November 2, 2023].
- Gaudreault, A. (1988). *Du littéraire au filmique: système du récit*. Paris: Méridiens Klincksieck (translated as GaudreaultA (2009)).
- Gaudreault, A. (2009). *From Plato to Lumière. Narration and Monstration in literature and cinema*. Toronto: University of Toronto Press.
- Gaudreault, A., and Gauthier, P. (2018). "Christian Metz, editing, and forms of alternation" in *Christian Metz and the codes of cinema. Film semiology and beyond*. eds. M. Tröhler, G. Kirsten and J. Zutavern (Amsterdam: Amsterdam University Press).
- Griffith, D.W. (1912). The Sunbeam. Available at: <https://archive.org/details/TheSunbeam> [Accessed November 2, 2023].
- Heath, S. (1973). Film/Cinetext/Text. *Screen* 14, 102–128. doi: 10.1093/screen/14.1-2.102
- ISO/IEC 8613-2 (1993). Information technology—open document architecture (ODA) and interchange format—document structures. ITU-T Recommendation T.412, Standard, International Telecommunication Union (ITU), Helsinki, 1993. Available at: <https://www.itu.int/rec/T-REC-T.412-199303-I/en> [Accessed November 2, 2023].
- Kawin, B. F. (1992). *How movies work*. Berkeley/Los Angeles/London: University of California Press.
- Metz, Christian (1968). *Essais sur la signification au cinéma, tome 1*. Paris: Klincksieck (translated as film language: A semiotics of the cinema by TaylorMichael.
- Metz, C. (1972). *Essais sur la signification au cinéma, tome 2*. Paris: Klincksieck.
- Metz, C. (1974a). *Film language: A semiotics of the cinema*. Chicago: The University of Chicago Press.
- Metz, Christian (1974b). *Language and cinema* (trans. by Umiker-SebeokDonna Jean) Paris: The Hague/Mouton.
- Middleton, Jim (2012). Deconstructing Griffith - A Girl and Her Trust (1912) prior to editing. Available at: <https://www.youtube.com/watch?v=BzGIETH-Olg> [Accessed March 16, 2024]
- Schlupkothen, F., and Schmidt, K.-H. (2022). "Legibility and Viewability. On the use of strict Incrementality in documents" in *Using documents: A multidisciplinary approach to document theory*. eds. G. Hartung, F. Schlupkothen and K.-H. Schmidt (Boston: De Gruyter. Berlin).
- Schmidt, K.-H. (2004). Zur chronologischen Syntagmatik von Bewegtbilddaten (II): Polyspatiale Alternanz. *Kodikas/Code* 27, 95–123.
- Schmidt, K.-H. (2008). Zur chronologischen Syntagmatik von Bewegtbilddaten (III): Deskriptive Syntagmen. *Kodikas/Code* 31, 137–189.
- Schmidt, K.-H., and Becher, M. (2017). Zur chronologischen Syntagmatik von Bewegtbilddaten (IV): Graduelle Handlungen und autonome Segmente. *Kodikas/Code*. 40, 239–277.
- Schmidt, K.-H., and Strauch, T. (2002). Zur chronologischen Syntagmatik von Bewegtbilddaten: Eine semiologische Reklassifikation der Syntagmatik von Metz (anhand einer Neuanalyse des Spielfilms *Adieu Philippe*). *Kodikas/Code* 25, 64–94.
- Stam, R., Burgoyne, R., and Flitterman-Lewis, S. (1992). *New vocabularies in film semiotics: Structuralism, post-structuralism and beyond*. London and NewYork: Routledge.
- TEI Consortium (2023). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.6.0. Last updated on 4th April 2023. TEI Consortium. Available at: <http://www.tei-c.org/Guidelines/P5/> [Accessed November 2, 2023].
- Thompson, K. (2011). A variation on a sunbeam: Exploring a Griffith Biograph film. Available at: <https://www.davidbordwell.net/blog/2011/09/05/a-variation-on-a-sunbeam-exploring-a-griffith-biograph-film/> [Accessed November 2, 2023].

## Author contributions

K-HS: Writing – original draft.

## Funding

The author declares that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



## OPEN ACCESS

## EDITED BY

Claudia Lehmann,  
University of Potsdam, Germany

## REVIEWED BY

Martin Luginbühl,  
University of Basel, Switzerland

## \*CORRESPONDENCE

Jana Pflaeging  
✉ jana.pflaeging@plus.ac.at

RECEIVED 19 December 2023

ACCEPTED 01 March 2024

PUBLISHED 20 May 2024

## CITATION

Pflaeging J (2024) Diachronic multimodality  
research – a mini-review.  
*Front. Commun.* 9:1358192.  
doi: 10.3389/fcomm.2024.1358192

## COPYRIGHT

© 2024 Pflaeging. This is an open-access  
article distributed under the terms of the  
[Creative Commons Attribution License](#)  
(CC BY). The use, distribution or reproduction  
in other forums is permitted, provided the  
original author(s) and the copyright owner(s)  
are credited and that the original publication  
in this journal is cited, in accordance with  
accepted academic practice. No use,  
distribution or reproduction is permitted  
which does not comply with these terms.

# Diachronic multimodality research – a mini-review

Jana Pflaeging\*

University of Salzburg, Salzburg, Austria

This *mini-review* gives an overview of *diachronic* multimodality research, an approach that multimodality scholars have pursued only rarely so far. Acknowledging the comparably large share of case studies in this line of research, this paper sets out by surveying empirical contributions. Many of them also provide discussions of selected theoretical and methodological aspects, which are subsequently collated and complemented with concepts from a number of more extensive theoretical proposals. Identifying main developments in diachronic multimodality studies and avenues for future work, this review seeks to support its growth into a mature research strand with a solid theoretical basis, a versatile methodological toolbox, and a broad range of research objects.

## KEYWORDS

diachronic, multimodality, development, stability, change, genre, review

## Introduction

Multimodality research has so far mainly been *synchronic* in nature (van Leeuwen, 2005), p. 26. The *development* of multimodal artefacts and interactions *over time* has received comparably little attention (Hiippala and Tseng, 2017; Stöckl, 2017, p. 263). At the same time, scholars have stressed the potential of adopting a *diachronic* perspective to gain a deeper understanding of the historical situatedness of multimodal communication (van Leeuwen, 2005, p. 142; Stöckl, 2009; Waller, 2012, p. 241; Bednarek and Caple, 2014, p. 150; Bateman et al., 2014, p. 10). This *mini-review* surveys existing empirical research with a diachronic angle and synthesizes previous theoretical and methodological proposals. It thereby seeks to contribute to the further growth of diachronic multimodality studies into a mature research strand.

## Previous empirical research

While *diachronic* multimodality research is generally scarce, a number of studies have appeared over the past 25 years that show a keen interest in the developments of multimodal communication and the factors that initiate and shape such processes (see contributions to special issues, e.g., Hiippala and Tseng, 2017, to collections, e.g., Hess-Lüttich et al., 1996; Holly, 1998; Schneider and Stöckl, 2011; Evangelista Allori et al., 2014; Hauser et al., 2014; Brock et al., 2019; Stöckl et al., 2019, and many stand-alone journal articles). Although points in time rarely overlap and periods differ in scope, most existing studies focus on *contemporary* multimodal practices and compare them to earlier related data. Perhaps not surprisingly, accounting for *change* (rather than *stability*) has been the most prominent research interest. The subsequent overview spotlights previous empirical contributions, showcases the diversity of approaches, scope of data-sets, and contributions to theory-building.



## Print media

A natural consequence, perhaps, of the long history of print media, the majority of diachronic multimodality research has focused on page-based static artefacts. Even before the study of multimodality gains traction in the mid-1990s (see, e.g., Kress and van Leeuwen, 1996; Yates and Orlikowski, 1992) provide a detailed account of the *business memo* and describe how multimodal features change over time. A further study of corporate discourse is Deng and Feng's (2022) work on the photographic representation of academics in 132 annual reports from six major universities in Hong Kong. Analysing materials from 1994/95–2015/16, they identify a development from context-rich group depictions to individual portraits that feature direct gaze.

Shifting the focus to mass-media journalism in the early 1990s, Bucher (1996, p. 35–41) and Bucher (1998, p. 70–94) describes an increase in the use of photographs and diagrams in German newspapers. The attested functional redistribution across modes is supported by a more modular page layout, which provides entry points to browsing readers (Bucher, 1996, p. 41–48). Durrani's (2020) study of 840 language-image combinations in *Time Asia* (1981–2010) describes a similar shift towards an *image-centric* news discourse. This development can equally be tracked in *National Geographic*, where feature articles (as published in 1915, 1965, and 2015) experience a reorganisation of the page space, with an increase in spreads that contain nothing but a large photograph and a short caption (Pflaeging, 2017a, 2017b). Spurred by changes in how captions are rhetorically tied to photographs, the double-page image-caption cluster develops into a key generic pattern that offers entry points, supports narrative interpretations across a longer feature article (Pflaeging, 2017a) and, over time, splits off into a stand-alone genre (Pflaeging, 2020a, p. 110–114). Stöckl's (2017) longitudinal study of the genre profiles, image types and language-image relations in the *MIT Technology Review* illustrates how changes at the genre-level are inextricably linked to higher-level developments, e.g., in genre profiles and in wider social domains (Stöckl, 2017, p. 273). Müller-Lancé's (2016, 2019) work on 400 French and German action-sports magazines (since the 1970s) identifies economic pressures as a main driving force of change.

Bateman (2008, p. 229–248; Bateman, et al. 2004, 2007 and Bateman, 2014a, p. 246–254) offers a diachronic account of expository discourse by the example of ornithological field guides (1924, 1972, 1994, 1996). Applying his *Genre-and-Multimodality* (GeM) framework, he tracks a shift towards increasingly complex layout structures (Bateman, 2008, p. 230–235) and a redistribution of content elements across a broader variety of semiotic modes (Bateman, 2008, p. 235–240). Focusing equally on educational materials, Bezemer and Kress (2009, 2016) study a corpus of 240 pages from 23 school textbooks (1930s, 1980s, 2000s). Reminiscent of Bateman's (2008) findings, they describe a trend towards a more extensive and complex use of page-based modes such as layout or typography (Bezemer and Kress, 2009, p. 256–260). The deployment of pictorial modes undergoes a shift from mainly photography (1930s, 1980s) to mostly drawings in the 2000s (Bezemer and Kress, 2009, p. 254). A recent contribution to this line of study is Keles and Yazan's (2021) critical discourse analysis of gender representations in a series of five school textbooks (1993–2019). Results shows that, rather than mitigating gender inequities, learning materials continue to consolidate heteronormative representational practices (Keles and Yazan, 2021, p. 138).

Scholars have also turned to the development of comics media. Cohn et al. (2017, p. 19), for instance, conducted a large-scale corpus analysis of American superhero comics (1940s–2010s) with a view to multimodal interactions, attentional framing, and semantic relationships between panels. Results suggest a trend towards more complex visual narrative structures and a reduced functional load carried by verbal elements (Cohn et al., 2017, p. 30, 34). Bateman et al. (2019, p. 216) investigate a similar data-set (1,260 pages from American superhero comics, 1940s–2000s), but place a more specific focus on page composition. Combining corpus-linguistic and multimodal methods, they offer a highly systematic medium-specific description of individual layout changes and correlations between them (Bateman et al., 2019, p. 216–223).

Hiippala's (2015) study applies the GeM-framework to a large set of tourist brochures and describes, for instance, a more elaborate use of layout after the advent of desktop-publishing. Turning to questions of genre ancestry, Molnar (2019) identifies the herbal, the trade card, and the private letter as predecessors of print-ad genres of the Enlightenment era. Further empirical work on advertisements offers theoretical remarks on why accelerated change is preferred in persuasive discourse to meet audience expectations (Cook, 2006, p. 224; Stöckl, 2010, 2014).

## Digital media

One of the earliest longitudinal studies of digital media is Zhang and O'Halloran's (2013) research on university homepages. Their work, informed by critical discourse analysis and social semiotics, reveals that education is increasingly framed as a lifestyle and that target audiences have shifted from national to global communities. Based on a corpus of 100 blogs, Schildhauer's (2016, Ch. 6) history of the *personal weblog* sheds light on the genre's development regarding layout, image types, and language-image links in response to technological advancements and emerging genre constraints. Pflaeging (2020b) and Dynel (2022) turn to viral online communication, studying soft-news items (listicles) and memes, respectively. While Dynel (2022, p. 73) finds that memes remained largely stable over 12 months, Pflaeging's (2020a, p. 240–242) data from 2014 and 2017 suggests that generic patterns may just as well bend quickly in order to persuade users to share a listicle with their social networks. Finally, Stamenković (2022) studies the stylistic development of video-game screens from the *Football Manager* series. He finds that, over time, screens feature less language and more pictorial elements and show a growing diversity of sub-canvases.

## Audio-Visual Media

While diachronic research on audio-visual media is comparably rare, Luginbühl (2014, 2019) provides a comprehensive, culture-contrastive account of TV news programs in the U.S. and Switzerland. Analysing 76 programs broadcast between 1949 and 2005, Luginbühl tracks changes in individual genres and the genre profile, and makes important contributions to a theory of genre development as well (Luginbühl, 2014, Ch. 9, 10; Luginbühl, 2015a; Luginbühl, 2019, p. 133–136). Brock (2019) pursues a similar interest in theory-building and describes various “genre-constitutive acts” of contemporary TV sitcoms (Brock, 2019, p. 123). Graakjær (2019), in

turn, investigates a corpus of 475 McDonald's TV commercials (2003–2018) focusing on the sonic logo. Inherent to musical practice, his study makes particularly clear that alteration and adaptation are an integral part of generic development (Graakjær, 2019, p. 580).

As suggested above, the vast majority of publications in diachronic multimodality research are empirical case studies, applying frameworks and concepts from social semiotics (e.g., Molnar 2019; Durrani, 2020; Deng and Feng, 2022), critical multimodal discourse analysis (e.g., Zhang and O'Halloran, 2013; Keles and Yazan, 2021; Dynel, 2022), empirically-oriented multimodal discourse analysis (as proposed by Bateman, Wildfeuer, and Hiippala, see e.g., Hiippala, 2015; Stamenković, 2022), and media/text linguistics (e.g., Bucher, 1996; Luginbühl, 2014; Schildhauer, 2016; Stöckl, 2017; Brock, 2019). Based on these empirical accounts, however, a number of theoretical assumptions can be deduced and synthesized with existing proposals.

## Theoretical aspects

### Genre as a focal point

For its strength in explaining the dynamics of stability and change of multimodal practices, many works in diachronic multimodality research have placed an emphasis on *genre*, usually with reference to paradigms such as “genre as social action” (Miller, 1984) or “genre as social semiotic” (Martin and Rose, 2008). Both traditions see communication as driven by social needs that arise from socio-historical contexts. As contexts change, participants explore new multimodal choices but naturally also rely on familiar patterns in text production and reception (Luginbühl, 2019, p. 132; also Eckkrammer, 2011, p. 196). Thus, while genres and patterns are “repurposed, redesigned and re-deployed” (Bateman et al., 2014, p. 10), the connection to genre ancestors is never lost (Lemke, 1999; also Molnar 2019; Sommer 2019). When genres *migrate* to new medial environments especially, established features are often initially retained for purposes of consolidation and orientation (Schildhauer, 2016; also Eckkrammer, 2011, p. 193–195).

Developments are stratified across various levels of abstraction (Bateman, 2008, p. 229; Luginbühl, 2014, 2019; Stöckl, 2017; Pflaeging, 2019, 2020a). At *pattern-level*, single (multi)modal choices (e.g., a layout pattern) may be perpetuated (*pattern stability*) or used more/less frequently over time (*pattern change* through *strengthening* or *weakening*), may be *internalised* through borrowing from other genres, or are novel creations (Pflaeging, 2017a, p. 259; Pflaeging, 2019, p. 77). Such individual developments were found to cluster and correlate at *genre-level* (see Cohn et al. 2017; Bateman et al., 2019; Deng and Feng, 2022). If the distribution of genre-typical patterns becomes increasingly untypical (Lemke, 1999), there is a strong indication of *genre change* (Luginbühl, 2014, p. 336), e.g., in the form of *standardisation* or *hybridisation* (Luginbühl, 2019, p. 136). When bundles of semiotic choices diverge noticeably from established use (Bhatia, 2014, p. 92), scholars speak of *genre split* (Lemke, 1999; Luginbühl, 2014, p. 335; Brock, 2015, p. 207; Schildhauer, 2016, p. 41; see Pflaeging, 2017a, p. 77–78). Developments like these, in turn, may result in changes at the *genre profile-level* (Luginbühl, 2015a, 2019, p. 133–134). By describing the *genre repertoire*, *genre frequencies* and *genre networks* typical of a given medium, statements can be made about a community's multimodal genre space (Bateman, 2008, p. 225–229).

## Factors that influence diachronic developments

Both empirical and theoretical works draw attention to the *factors* that shape the development of multimodal practices. Below, such factors will be discussed individually, although they are tightly intertwined (Pflaeging, 2017a,b; Cohn et al. 2017, p. 32; Stöckl, 2017; Graakjær, 2019, p. 573).

Among the most central driving forces is communicative function (Brock, 2019, p. 120; also Eckkrammer, 2011, p. 196; Stöckl, 2017, p. 273), which is most apparent when *genre* is attributed a key role in the description of communicative practice. In advertising (Cook, 2006; Molnar, 2019) and viral online discourse especially (Pflaeging, 2020a), where texts are meant to persuade to buy or click, this may result in accelerated change. Brock (2019) makes a similar argument for humour, where textual patterns quickly fail to evoke humorous interpretations. The functional range of a genre may also shift or become more diverse over time, which equally has repercussions on the text design (Brock, 2019, p. 120).

Genre expectations are another factor of influence (Hiippala and Tseng, 2017; Pflaeging, 2017a; Molnar 2019, p. 31). As socio-cognitive entities, genres are “devices for sense-making” (Lomborg, 2014, p. 45) and serve as frames of reference for the interpretation of multimodal artefacts and interactions (see also Cohn et al. 2017, p. 33). This creates the need for *perpetuating* established textual patterns to some extent (Luginbühl, 2019, p. 130, in ref. to Schildhauer, 2016, p. 259–261), especially in mass-media journalism. Here, participants experience a *disjunction of place and time* between text production and reception (Pflaeging, 2017a, p. 257–258), audiences remain imaginary and feedback is naturally “delayed” (Bell, 1984, 1991a, 1991b). While journalistic practice is equally subject to innovation, diachronic studies have revealed tendencies towards *standardisation* (Pflaeging, 2017b, p. 198; Luginbühl, 2019, p. 135; Pflaeging, 2020a, p. Fn. 7), using corporate-design manuals and templates. Advertisements, viral soft-news items, comics or sitcoms, on the other hand, can be *expected* to exploit the potential for creativity more. Finally, developments may also be shaped by recipients' awareness of communicative trends in the media landscape more generally, e.g., the rise of Japanese mangas on the American market (Cohn et al. 2017, p. 24), or a growing *visualisation* (Bucher, 1996; Caple, 2013, p. 7; also Krotz, 2015), *clusterisation* (Bucher, 1996), or *atomization* (Knox, 2007, p. 48) in journalism.

In advertising (Stöckl, 2010, p. 150), humorous discourse (Brock, 2019, p. 122), journalism (Pflaeging, 2017b, p. 201–202), and even more so in comics (Cohn et al. 2017, p. 32) and social media discourse (Dynel, 2022), change in multimodal practices may be spurred by an individual's creative ways of breaking with conventions, which may be deliberate or entirely coincidental (Pflaeging, 2017b, p. 201–202; Brock, 2019, p. 123). Change and stability may also result from developments in participant constellations more generally. Cohn et al. (2017, p. 32), for instance, assume that a revaluation of the role of the penciler may have led to the rise of *visual* storytelling in comics. In a similar vein, entire production teams may undergo processes of expansion and professionalisation, e.g., at *National Geographic* (Pflaeging, 2017b, p. 197–198; also Stöckl, 2017, p. 263). Developments may equally be fueled by changes in audiences, which may grow, diversify (Pflaeging, 2017a, p. 197–198), or experience new degrees of

involvement (see Brock, 2019, p. 114; Sommer 2019, p. 246; see also Meier and Marx 2019).

Medium-related advancements are also known to drive the development of multimodal practices, although they never *determine* them (Luginbühl, 2015b, p. 14; Pflaeging, 2017b, p. 192; Stöckl, 2017, p. 263; Luginbühl, 2019, p. 136; Brock, 2019, p. 120). Affordances of a given medium (and its material, technological and infrastructural qualities, Pflaeging, 2017b, p. 202; Stöckl, 2017, p. 273) may limit communicative choices, but participants are often seen to overcome supposed restrictions (Luginbühl, 2015b, p. 14–15). Shifting the focus further to the sociological dimension of media, practices may be influenced by a growing *mediatization* of social spheres (Androutsopoulos, 2014), incl. Journalism (Kammer, 2013) where multimodal practices have become reminiscent of a so-called *media logic* (Altheide and Snow, 1992; Stöckl, 2017, p. 272–273).

Such developments are closely tied to socio-economic factors as well (Eckkrämmer, 2011, p. 191). Müller-Lancé (2016, p. 596) study of action-sports magazines, for instance, illustrates how tight budgets (due to changing sales figures and sponsorships) may result in an outsourcing of journalistic tasks to guest contributors and even audience members. Likewise, early journalistic practice at *National Geographic* was shaped by limited funds for photographic reproduction and staff. A steep rise in circulation numbers soon led to a much healthier financial situation, more editorial staff, more advanced production techniques, and, ultimately, changes at the discourse-level (Pflaeging, 2017b). Though cause-effect relations seem less immediate, communicative developments can also be impacted by the socio-political dynamics in a given cultural and historical context (Brock, 2019, p. 121; Luginbühl, 2019). Factors such as the rise of *neoliberalism* and the *marketisation* and *commodification* of higher-education are typically identified within critical multimodal discourse analysis, where systemic-functional-linguistic approaches allow for constructing a connection between texts and socio-cultural contexts (Zhang and O'Halloran, 2013; Deng and Feng, 2022).

## Data collection and methodological aspects

### Data collection and corpus compilation

Diachronic multimodality research requires data from at least two different points in time, although choosing them is far from trivial. To cover a longer phase of continuous journalistic output, Luginbühl (2014) and Pflaeging (2017b, 2019) sample from (evenly distributed) points in time, with smaller and larger gaps in between (also Stöckl, 2017; Sommer 2019; Pflaeging, 2020b; Stamenković, 2022). Their approaches differ, however, in that Luginbühl's collection was guided by news events, whereas Pflaeging accounted for a phase of genre consolidation after moments of editorial reorientation. Stöckl (2017, p. 274) and Brock (2019), in turn, propose to focus on texts that capture moments of innovation in a genre's or medium's history. Yet other studies seek to reflect more closely the amount of available data, e.g., by sampling from densely-distributed points in time (Molnar 2019; Dynel, 2022) or by selecting (nearly) all data available for a given period, medium, or genre (e.g., Hiippala, 2015; Graakjær, 2019; Keles and Yazan, 2021; Deng and Feng, 2022). Studies on web-based genres have used the *Internet Archive*

(Schildhauer, 2016, p. 218) and web-scraping techniques for data collection (Dynel, 2022).

As results are to be related or relatable to map a development, diachronic data-sets need to show similarities in cultural contexts (e.g., Luginbühl, 2014; Sommer 2019), medium (e.g., Luginbühl, 2014; Keles and Yazan, 2021; Deng and Feng, 2022), genre (e.g., Schildhauer, 2016; Pflaeging, 2020a, 2020b; Dynel, 2022; Stamenković, 2022) or, at least, a strong similarity in textual function (Eckkrämmer, 2011, p. 203), topic (Stöckl, 2017, p. 263), or mode (Graakjær, 2019; Bateman et al., 2019; Deng and Feng, 2022). Some studies have explored an ethno-categorical approach to genre in order to ensure comparability (Luginbühl, 2014; Schildhauer, 2016; Pflaeging, 2019; Sommer 2019) and to avoid circularity in sampling (Thomas, 2019, p. 86). When diachronic corpora are meant to cover extensive periods of time, their compilation can be challenging due to a limited availability of materials or semiotically authentic, objective accounts (e.g., only a written report of a dynamic theatrical performance), or an insufficient preservation of historical materials that requires time-intensive digitalisation. Also, adding rich meta-data that is true to a given socio-historical context is not always straightforward (e.g., allocating genre categories, authorship, dates of publication) and working with historical materials generally requires a change of perspective.

### Accounting for different socio-historical contexts

Pragmatics- and discourse-oriented approaches to multimodal analysis (Bateman, 2014b, Ch. 11–12) emphasise the sensitivity of discourse interpretations to contexts of use. The analysis of multimodal texts from sometimes distant points in time thus requires an awareness and careful reconstruction of socio-historical contexts (see, e.g., Bateman, 2008, p. 237; Luginbühl, 2014; Schildhauer, 2016; Pflaeging, 2017b; Molnar 2019; Brock, 2019; Cohn et al. 2017, p. 33–34). Striking a balance between detailed, context-sensitive descriptions of qualitative differences and the aim to chart general developments remains a challenge in diachronic multimodality research.

### Analysing large(r) corpora and open research data

Empirical multimodality research has recently begun to turn to larger-scale data-sets (Pflaeging et al., 2021, p. 13) and research objects of considerable modal complexity (Bateman, 2021, p. 35). With diachronic corpora potentially multiplying the data under study, the issue of feasibility (Hiippala, 2017, p. 277, also in ref. to Thomas, 2009, p. 245) has become pressing. To scale-up, scholars explore new ways of annotating data using computational (Hiippala, 2016) and crowd-sourcing methods (Hiippala, 2023), and to implement statistics (Hiippala, 2015; Cohn et al. 2017; Bateman et al., 2019; also Bateman et al., 2017, Ch. 6; Bateman and Hiippala, 2020) and AI (Semedo and Magalhães, 2019).

The increasing scale of available data-sets is not least due to recent efforts in the spirit of the open-research-data paradigm (see the FAIR guidelines, Wilkinson et al., 2016 as well as an adapted version for cultural heritage objects, Koster and Woutersen-Windhouwer, 2018). Although this trend poses challenges to scholars seeking to compile historical corpora (e.g., due to time-intensive digitalisation for it to



be machine-readable and a careful consideration of copyrights), such efforts are necessary to make diachronic data-sets available to various communities of researchers.

## Discussion and outlook

Despite its scarcity, previous diachronic multimodality research has shown that our understanding of communicative practices today and our predictions of the future benefit from a close consideration of the past. Advances in theory-building suggest that developments in multimodal artefacts and interactions are driven by a broad range of interrelated factors and play out at various levels of the discourse. Solid theoretical frameworks are now in place (e.g., Bateman, 2008; Bateman et al., 2017; Wildfeuer et al., 2020) to systematically trace phenomena of stability and change across extensive periods of time. Future research will likely intensify the use of (semi-) automated and statistical methods to process large-scale multimodal corpora, compute correlations between changing patterns, and output significant factors of influence. The rise of digital archives and web-scraping methods opens up whole new worlds of multimodal data yet to be explored. All of these aspects make *diachronic multimodality research* a promising and timely endeavour: as with multimodality research in general, “there are more open questions than there are answers” (Bateman et al., 2017, p. 123) and “remember, there is always the ‘next’ paper :-)” (Bateman, 2020, pers. correspondence).

## Author contributions

JP: Writing – original draft, Writing – review & editing.

## References

- Altheide, D. L., and Snow, R. (1992). Media logic and culture: a reply to Oakes. *Int. J. Polit. Cult. Soc.* 5, 465–472. doi: 10.1007/BF01423902
- Androutsopoulos, J.K., ed. (2014). *Mediatization and sociolinguistic change*. Berlin: de Gruyter
- Bateman, J.A. (2008). *Genre and multimodality. A Foundation for the Systematic Analysis of multimodal documents*. Palgrave Macmillan: London
- Bateman, J.A. (2014a). Genre in the age of multimodality: some conceptual refinements for practical analysis. *Evolution in genre: Emergence, variation, multimodality*, (Eds.) P. Evangelisti Allori, J.A. Bateman and V.K. Bhatia, Bern: Peter Lang, 237–269
- Bateman, J.A. (2014b). *Text and image: a critical introduction to the visual/verbal divide*. London/New York: Routledge
- Bateman, J.A. (2020). Personal correspondence via Skype. 17.10.2020, 15:52
- Bateman, J. A., Veloso, F. O., and Lau, Y. L. (2019). On track of visual style: a diachronic study of page composition in comics and its functional motivation. *Visual Communication* 20, 209–247. doi: 10.1177/1470357219839101
- Bateman, J.A. (2021). Dimensions of materiality: towards an external language of description for empirical multimodality research. *Empirical multimodality research*, (Eds.) J. Pflaeging, J. Wildfeuer and J.A. Bateman, Berlin/Boston: de Gruyter, 35–63
- Bateman, J.A., Delin, J.L., and Henschel, R. (2004). Multimodality and empiricism: preparing for a corpus-based approach to the study of multimodal meaning-making. *Perspectives on multimodality*, (Eds.) E. Ventola, C. Charles and M. Kaltenbacher, Amsterdam: Benjamins, 65–89
- Bateman, J.A., Delin, J.L., and Henschel, R. (2007). Mapping the multimodal genres of traditional and electronic newspapers. *New directions in the analysis of multimodal discourse*, (Eds.) T.D. Royce and W.L. Bowcher, Mahwah, NJ: Lawrence Erlbaum, 147–172
- Bateman, J.A., Evangelisti Allori, P., and Bhatia, V.K. (2014). Evolution in genre: emergence, variation, multimodality. *Evolution in genre*, (Eds.) P. Evangelisti Allori, J.A. Bateman and V.K. Bhatia, Bern: Peter Lang, 9–16
- Bateman, J. A., and Hiippala, T. (2020). Statistics for Multimodality: why, when, how - an invitation. doi: 10.31235/osf.io/7j3np
- Bateman, J.A., Wildfeuer, J., and Hiippala, T. (2017). *Multimodality: foundations, research and analysis – a problem-oriented introduction*. Berlin/Boston: de Gruyter
- Bednarek, M., and Caple, H. (2014). Why to news values matter? Towards a new methodological framework for analysing news discourse in critical discourse analysis and beyond. *Discourse Soc.* 25, 135–158. doi: 10.1177/0957926513516041
- Bell, A. (1984). Language style as audience design. *Lang. Soc.* 13, 145–204. doi: 10.1017/S004740450001037X
- Bell, A. (1991a). Audience accommodation in the mass media, *Contexts of accommodation: developments in applied linguistics*, (Eds.) H. Giles, J. Coupland and N. Coupland, Cambridge: Cambridge University Press, 69–102
- Bell, A. (1991b). *The language of news media*. Oxford: Blackwell
- Bezemer, J., and Kress, G. (2009). Visualizing English: a social semiotic history of a school subject. *Vis. Commun.* 8, 247–262. doi: 10.1177/1470357209106467
- Bezemer, J., and Kress, G. (2016). The textbook in a changing multimodal landscape. *Handbuch Sprache im Multimodalen Kontext*, (Eds.) N.-M. Klug and H. Stöckl, Berlin/Boston: de Gruyter, 476–498
- Bhatia, V.K. (2014). *Worlds of written discourse: a genre-based view*. London: Bloomsbury
- Brock, A. (2015). Comedy Panel Show, Dramey und Improv-Comedy: Zur kulturellen Ausdifferenzierung komischer Fernsehgenres in Großbritannien. *Das Komische in der Kultur*, (Eds.) S. Neuhaus, H. Diekmannshenke und U. Schaffers, Marburg: Tectum, 193–208

## Funding

The author declares financial support was received for the research, authorship, and/or publication of this article. This work was supported by the University of Salzburg Publication Fund.

## Acknowledgments

I would like to thank Martin Luginbühl for his insightful comments on an earlier version of this manuscript. I am also grateful to Claudia Lehmann, Janina Wildfeuer und Tamara Drummond for their efforts in editing this special issue.

This paper is dedicated to John A. Bateman, who has been a major driving force in the development of our research field and an invaluable guide in my own exploration of multimodality.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



- Brock, A. (2019). The emergence of contemporary British TV sitcoms, *Genre emergence: developments in print, TV and digital media*, (Eds.) A. Brock, J. Pflaeging and P. Schildhauer, Frankfurt am Main: Peter Lang, 107–127
- Brock, A., Pflaeging, J., and Schildhauer, P., (Eds.) (2019). *Genre emergence: developments in print, TV and digital media*. Frankfurt am Main: Peter Lang
- Bucher, H.J. (1996). Textdesign-Zaubermittel der Verständlichkeit? Die Tageszeitung auf dem Weg zum interaktiven medium, *Textstrukturen im Medienwandel*, (Eds.) E.W.B. Hess-Lüttich, W. Holly and U. Püschel, Frankfurt am Main: Peter Lang, 31–60
- Bucher, H.J. (1998). Vom Textdesign zum Hypertext: Gedruckte und elektronische Zeitungen als nicht-lineare Medien, *Medien im Wandel*, (Ed.) W. Holly, Opladen/Wiesbaden: Westdeutscher Verlag GmbH, 63–102
- Caple, H. (2013). *Photojournalism: a social semiotic approach*. Basingstoke: Palgrave Macmillan
- Cohn, N., Taylor, R., and Pederson, K. (2017). A picture is worth more words over time: multimodality and narrative structure across eight decades of American superhero comics. *Multimodal Communication*. 6, 19–37. doi: 10.1515/mc-2017-0003
- Cook, G. (2006). *The discourse of advertising*. London/New York: Routledge.
- Deng, Y., and Feng, D. (2022). From researchers to academic entrepreneurs: a diachronic analysis of the visual representation of academics in university annual reports. *Vis. Commun.* 80:147035722211021. doi: 10.1177/14703572221102180
- Durrani, S. (2020). Previewing news stories: how contextual cohesion contributes to the creation of news stories, *Shifts toward image-centricity in contemporary multimodal practices*, (Eds.) H. Stöckl, H. Caple and J. Pflaeging, New York/London: Routledge, 123–145
- Dynel, M. (2022). The life of COVID-19 mask memes: a diachronic study of the pandemic memescape. *Communicar.* 30, 73–85. doi: 10.3916/C72-2022-06
- Eckkrämmer, E.M. (2011). Diachrone Medienanalyse: Zur Analyse multimodaler Vertextungsstrategien in historischer Sicht, *Medientheorien und Multimodalität: Ein TV-Werbespot – Sieben methodische Beschreibungsansätze*, (Eds.) J.G. Schneider and H. Stöckl, Köln: Herbert von Halem, 190–215
- Evangelisti Allori, P., Bateman, J.A., and Bhatia, V.K., (Eds.) (2014). *Evolution in genre: emergence, variation, multimodality*. Bern: Peter Lang
- Graakjær, N. J. (2019). Sounding out 'i'm lovin' it: a multimodal discourse analysis of the sonic logo in commercials for McDonald's 2003–2018. *Crit. Discourse Stud.* 16, 569–582. doi: 10.1080/17405904.2019.1624184
- Hauser, S., Kleinberger, U., and Roth, K.S. (Eds.) (2014). *Musterwandel-Sortenwandel: aktuelle tendenzen der diachronen Text(sorten)linguistik*. Bern: Peter Lang
- Hess-Lüttich, E.W.B., Holly, W., and Püschel, U. (Eds.) (1996). *Textstrukturen im Medienwandel*. Frankfurt am Main: Peter Lang
- Hiippala, T. (2015). *The structure of multimodal documents: an empirical approach*. London/New York: Routledge
- Hiippala, T. (2016). Semi-automated annotation of page-based documents within the genre and multimodality framework, Proceedings of the 10th SIGHUM workshop on language Technology for Cultural Heritage, social sciences, and humanities, Berlin: Association for Computational Linguistics, 84–89
- Hiippala, T. (2017). An overview of research within the genre and multimodality framework. *Discourse, Context & Media*. 20, 276–284. doi: 10.1016/j.dcm.2017.05.004
- Hiippala, T. (2023). Corpus-based insights into multimodality and genre in primary school science diagrams. *Vis. Commun.* 29:147035722311618. doi: 10.1177/14703572231161829
- Hiippala, T., and Tseng, C.-I. (2017). Editorial: media evolution and genre expectations. *Discourse, Context & Media*. 20, 157–159. doi: 10.1016/j.dcm.2017.11.001
- Holly, W., (Ed.) (1998). *Medien im Wandel*. Opladen/Wiesbaden: Westdeutscher Verlag GmbH
- Kammer, A. (2013). The mediatization of journalism. *MediaKultur*. 29, 141–158. doi: 10.7146/mediekultur.v29i54.17385
- Keles, U., and Yazan, B. (2021). Gender in new headway: a diachronic, multimodal critical discourse analysis. *Lang. Teach.* 54, 132–138. doi: 10.1017/S0261444820000579
- Knox, J. (2007). Visual-verbal communication on online newspaper home pages. *Vis. Commun.* 6, 19–53. doi: 10.1177/1470357207071464
- Koster, L., and Woutersen-Windhauer, S. (2018). FAIR principles for library, archive and museum collections: a proposal for standards for reusable collections. *code{4}lib Journal*. 40
- Kress, G., and van Leeuwen, T. (1996). *Reading images: the grammar of visual design*. London/New York: Routledge
- Krotz, F. (2015). Mediatisierung und die wachsende Bedeutung visueller Kultur: Zum Verhältnis zweier Kommunikationswissenschaftlicher Metaprozesse, *Visualisierung-Mediatisierung: Bildliche Kommunikation und bildliches Handeln in mediatisierten Gesellschaften*, (Eds.) K. Lobinger and S. Geise, Köln: Halem Verlag, 18–36
- Lemke, J.L. (1999). Typology, topology, topography: Genre semantics, Available at: <http://academic.brooklyn.cuny.edu/education/jlemke/papers/Genre-topology-revised.htm>, (Accessed Dec 18, 2023).
- Lomborg, S. (2014). *Social media, social genres: making sense of the ordinary*. London/New York: Routledge
- Luginbühl, M. (2014). *Medienkultur und Medienlinguistik: Komparative Textsortengeschichte(n) der amerikanischen 'CBS Evening News' und der Schweizer 'Tagesschau'*. Bern: Peter Lang
- Luginbühl, M. (2015a). Genre profiles as intermediate analytical level for cultural genre analysis, *Genre studies around the globe*, (Eds.) N. Artemeva and A. Freedman, Bloomington: Inkshed Publications, 251–273
- Luginbühl, M. (2015b). On mediality and culturality. *10plus1: Living Linguistics*. 1, 9–26
- Luginbühl, M. (2019). Genre emergence and change as indicator and origin of cultural change, *Genre emergence: developments in print, TV and digital media*, (Eds.) A. Brock, J. Pflaeging and P. Schildhauer, Frankfurt am Main: Peter Lang, 129–159
- Martin, J.R., and Rose, D. (2008). *Genre relations: mapping culture*. London: Equinox
- Meier, S., and Marx, K. (2019). Doing genre in the digital media, *Genre emergence: developments in print, TV and digital media*, (Eds.) A. Brock, J. Pflaeging and P. Schildhauer, Frankfurt am Main: Peter Lang, 191–212
- Molnar, S. (2019). The birth of the print ad genre, *Genre emergence: developments in print, TV and digital media*, (Eds.) A. Brock, J. Pflaeging and P. Schildhauer, Frankfurt am Main: Peter Lang, 29–50
- Miller, C. R. (1984). Genre as social action. *Q. J. Speech.* 70, 151–167. doi: 10.1080/00335638409383686
- Müller-Lancé, J. (2016). *Trendsportmagazine in Deutschland und Frankreich: Eine medienlinguistische Analyse*. Landau: Verlag Empirische Pädagogik
- Müller-Lancé, J. (2019). The development of genres in German and French action sports magazines: how economic interests affect text types, *Genre emergence: developments in print, TV and digital media*, (Eds.) A. Brock, J. Pflaeging and P. Schildhauer, Frankfurt am Main: Peter Lang, 51–72
- Pflaeging, J. (2017a). Tracing the narrativity of National Geographic feature articles in the light of evolving media landscapes. *Discourse, Context & Media*. 20, 248–261. doi: 10.1016/j.dcm.2017.07.003
- Pflaeging, J. (2017b). Changing potentials and their use: the case of popular science journalism, *Communication forms and communicative practices: new perspectives on communication forms, affordances and what users make of them*, (Eds.) A. Brock and P. Schildhauer, Frankfurt am Main: Peter Lang, 181–208
- Pflaeging, J. (2019). Beyond genre names: diachronic perspectives on genre indexation in print magazines, *Genre emergence: developments in print, TV and digital media*, (Eds.) A. Brock, J. Pflaeging and P. Schildhauer, Frankfurt am Main: Peter Lang, 73–104
- Pflaeging, J. (2020a). On the emergence of image-centric popular science stories in National Geographic, *Shifts toward image-centricity in contemporary multimodal practices*, eds. H. Stöckl, H. Caple and J. Pflaeging, London/New York: Routledge, 97–122
- Pflaeging, J. (2020b). Diachronic perspectives on viral online genres: from images to words, from lists to stories, *Visualizing digital discourse: interactional, institutional and ideological perspectives*, (Eds.) C. Thurlow, C. Dürscheid and F. Diémoz, Berlin/Boston: de Gruyter, 227–244
- Pflaeging, J., Bateman, J.A., and Wildfeuer, J. (2021). Empirical multimodality research: the state of play, *Empirical multimodality research*, (Eds.) J. Pflaeging, J. Wildfeuer and J. A. Bateman, Berlin/Boston: de Gruyter, 3–32
- Schildhauer, P. (2016). *The personal weblog: a linguistic history*. Frankfurt am Main: Peter Lang
- Schneider, J.G., and Stöckl, H., eds. (2011). *Medientheorien und Multimodalität: Ein TV-Werbespot-Sieben methodische Beschreibungsansätze*. Köln: Herbert von Halem
- Semedo, D., and Magalhães, J. (2019). Diachronic cross-modal embeddings, Proceedings of the 27th ACM international conference on multimedia, 2061–2069
- Sommer, J. M. (2019). Emergence of online comments in popular science discourse, *Genre emergence: developments in print, TV and digital media*, (Eds.) A. Brock, J. Pflaeging and P. Schildhauer, Frankfurt am Main: Peter Lang, 235–259
- Stamenković, D. (2022). The stylistic journey of a video game: a diachronic approach to multimodality in the football manager series, *Stylistic approaches to pop culture*, (Eds.) C. Schubert and V. Werner, London/New York: Routledge, 227–246
- Stöckl, H. (2009). The language-image-text: theoretical and analytical inroads into semiotic complexity. *AAA – Arbeiten aus Anglistik und Amerikanistik*. 34, 3–28
- Stöckl, H. (2010). Textsortenentwicklung und Textverstehen als Metamorphosen: Am Beispiel der Werbung, *Mediale transkodierungen: metamorphosen zwischen Sprache, Bild und Ton*, (Ed.) H. Stöckl, Heidelberg: Universitätsverlag Winter, 145–172
- Stöckl, H. (2014). 'He begs to inform every person interested': a diachronic study of address and interaction in print advertising. *Anglistik Inte. J. Engl. Stud.* 25, 81–106

- Stöckl, H. (2017). Multimodality in a diachronic light: tracking changes in text-image relations within the genre profile of the MIT technology review. *Discourse, Context & Media*. 20, 262–275. doi: 10.1016/j.dcm.2017.07.001
- Stöckl, H., Caple, H., and Pflaeging, J., (Eds.) (2019). *Shifts toward image-centricity in contemporary multimodal practices*. London/New York: Routledge
- Thomas, M. (2009). *Localizing pack messages: a framework for corpus-based cross-cultural multimodal analysis*. PhD Thesis. Leeds: University of Leeds
- Thomas, M. (2019). Making a virtue of material values: tactical and strategic benefits for scaling multimodal analysis, *Multimodality: towards a new discipline*, (Eds.) J. Wildfeuer et al., Berlin/Boston: de Gruyter, 69–91
- van Leeuwen, T. (2005). *Introducing social semiotics*. London/New York: Routledge
- Waller, R. (2012). Graphic literacies for a digital age: the survival of layout. *Inf. Soc.* 28, 236–252. doi: 10.1080/01972243.2012.689609
- Wildfeuer, J., Bateman, J.A., and Hiippala, T. (2020). *Multimodalität: Grundlagen, Forschung und Analyse – Eine problemorientierte Einführung*. Berlin/Boston: de Gruyter
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*. 3, 1–9. doi: 10.1038/sdata.2016.18
- Yates, J., and Orlikowski, W. J. (1992). Genres of organizational communication: a structural approach to studying communication and media. *Acad. Manag. Rev.* 17, 299–326. doi: 10.2307/258774
- Zhang, Y., and O'Halloran, K. L. (2013). 'Toward a global knowledge enterprise': university websites as portals to the ongoing marketization of higher education. *Crit. Discourse Stud.* 10, 468–485. doi: 10.1080/17405904.2013.813777



## OPEN ACCESS

## EDITED BY

Claudia Lehmann,  
University of Potsdam, Germany

## REVIEWED BY

Alexander Bergs,  
Osnabrück University, Germany  
Andreas Rothenhöfer,  
University of Bremen, Germany  
Gabriele Marino,  
University of Turin, Italy

## \*CORRESPONDENCE

Wolfgang Wildgen  
✉ wildgen@uni-bremen.de

RECEIVED 07 December 2023

ACCEPTED 23 April 2024

PUBLISHED 21 May 2024

## CITATION

Wildgen W (2024) The cognitive roots of multimodal symbolic forms with an analysis of multimodality in movies.  
*Front. Commun.* 9:1352252.  
doi: 10.3389/fcomm.2024.1352252

## COPYRIGHT

© 2024 Wildgen. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# The cognitive roots of multimodal symbolic forms with an analysis of multimodality in movies

Wolfgang Wildgen\*

Institut für Allgemeine und Angewandte Sprachwissenschaft (IAAS), Sprach- und Literaturwissenschaften, Universität Bremen, Bremen, Germany

Condillac's (1754) "Traité des sensations" is the philosophical background of modern discussions on the relationship between perception and multimodal communication. The differences between perception and communication and the transitions between them are discussed with a focus on odor and color. It becomes clear that even at this primary level, the complex interactions of different modalities are the precondition for effective and rich communication. The second part discusses Cassirer's "Philosophy of Symbolic Forms" as a relevant framework for multimodality studies. Basic aspects are first commented on with a focus on music and visual art. The interaction is even more complex and rich in the case of language; the difficulty of large symbolic forms is mainly due to semantic composition and only to a lesser degree to syntactic concatenation. The first must merge/blend different semantic spaces. It must allow for the plurality of levels of integration from the lexical level, the level of phrases and sentences, up to texts and discourse. The third part focuses on multimodality in film. It treats the representation of movement and action in (film) narratives, the visual perception and representation/communication of movement and action, and the integration of music, moving images, and language.

## KEYWORDS

multimodality, symbolic forms, cognitive roots, music, visual artifacts, language, movies, semiotics

## 1 Sensation, knowledge, and modes of communication

The relationship between sensation and knowledge was the focus of the "Traité des Sensations" ("Treatise on Sensations"; Condillac 1754/1970). It presents the following thought experiment. A marble statue is successively endowed with sensuous experience. Condillac begins with olfaction. The statue first is what it smells; second, it distinguishes pleasant and unpleasant odors. Finally, comparison and memory select and stabilize the attention movement after the stimulus' reception. The difference is passive if motivated by a stimulus and active if motivated by memory (cf. Condillac, 1754/1970: 49). With the two principles of pleasure/aversion and memory/comparison, Condillac treats hearing, taste, and finally, sight, consisting of light and color. He suggests a system of ideas based on olfaction. In his earlier treatise "Essai Sur l'Origine des Connaissances Humaines" (An Essay on the Origin of Human Knowledge; 1746/1982), he tried to reduce the theory put forward by Locke (1690/1975) in his "An Essay Concerning Human Understanding" to one principle (sensation) and to explain the rise of human language and culture.

From the present perspective, the "sensualistic" position of Condillac was a sound synthesis of ideas put forward in the discussions of modernity (sixteenth to eighteenth

century) but underestimated the problem's complexity. The composition of larger structures in visual artifacts, musical composition, and text/literature asks for large-scale organizational principles, which may be called the "architecture" of human communication.

## 1.1 The complexity of symbolic forms beyond the senses and their integration

Despite all advances, the following significant problems could not be solved:

- How are meanings grounded in our (subconscious) sensory activities, which are adapted by evolution to selected features of our environment? The quality of this grounding (the "correct" selection, the stable transfer of relevant structural relations up to the highest levels of cognition) is crucial for the functional fitness of the "animal symbolicum," as humans are called by Cassirer (1995).
- How can individual thinking (as representation) be economical and still represent external reality?

Condillac considered perception as the first semiotic level and signs/semiotic entities necessary to stabilize and organize memory and thinking as a further one.<sup>1</sup> However, this scale of an unfolding of human communication is incomplete because a fundamental conflict exists between individual perception/action and social communication. The first field has the individual brain (including the sensory organs) and the individual development (maturation and essential learning in a suitable environment) as its domain. The second field concerns the adaptation to a culturally transmitted system of linguistic rules, social beliefs, knowledge, social action, and perception.

Although learning and socialization link both domains, the fundamental mechanisms differ. The organization of the brain cannot be identified with the organization of a community, and communication underlies other conditions of coding, transmission, functional adequacy, and economy than that operative in sensory organs and cortical centers of perception and motor planning. The differences become relevant in the case of olfaction and color and, thus, the contribution of sensory experience to unfolding symbolic forms.<sup>2</sup> Therefore, we shall consider this critical transition decisive for

evolution of human societies as a first step.<sup>3</sup> In the second step, we analyze complex communication and multimodal symbolic forms. Eventually, multimodal communication is studied in the interaction of visual, textual, and musical communication in movies.

## 1.2 The mode of communication about odors

If both the speaker and the audience are exposed to the same smell, it is enough to call the audience's attention to the scent and then perform a speech act referring to the (supposed) common perception.<sup>4</sup> It is only if the odor is not present and must be represented that the problem of proper categorization and characterization occurs.<sup>5</sup> It presupposes a stable system of meanings grounded in similar olfactory experiences. In a community where this demand does not occur frequently, communication regarding odors will be very insecure, unstable, poor, or meaningless. Thus, in the extreme case, every participant in a conversation may be perfectly able to perceive and, if necessary, distinguish a large set of odors but unable to communicate these distinctions. The required ability concerns first the stable grounding in personal perceptual experience, second efficient coordination with the perceptual patterns in those with whom one wants to communicate, and third, the invention and use of labels (lexical or periphrastic). Thus, it does not help to learn words for

3 The evolutionary sequence where basic capacities of perception and motor control were developed (e.g., in the Cambrian revolution) reappears in the rooting of higher mental capacities (e.g., language and other symbolic forms) in perception and motor control.

4 The notions "mode" and "modality" used in this contribution always imply communication; "modality" is an abstraction regarding modes of communication. In the present study, communication is primarily human communication, not between animals, plants, cells, or bodily substances, as in zoosemiotics. Cassirer's term "symbolic forms" is more specific insofar as he assumes that the symbolic capacity is a defining feature of humans. Since the first half of the last century, when Cassirer developed his "Philosophy of Symbolic Forms," our knowledge of the evolution of primates and humans was further advanced; therefore, hypotheses assuming the isolation of human symbolic capacity must be revised. In our context, communication is restricted to humans, and the question of generalizing our analysis beyond humans is not touched. The list of "modes" varies with the authors. Kress (2010:79) mentions: "image, writing, layout, gesture, speech, moving image, soundtrack." He refers to "meaning-making" as a common feature. The crux with "meaning making" as the central criterion is that the traditional notion of "meaning" in linguistics does not match parallel features called signification, significance, or functional effect in painting, architecture, music, and all modalities different from language. In this essay, we distinguish primary modes (of communication) linked to human senses: touch, odor, sight, and hearing. Consequently, language is considered a secondary mode, mixing different sense-related modes or being dissociated from these modes. Speech and writing would be modes dominated by the linguistic mode.

5 Sections 1.2 and 1.3 use materials written by the author to prepare the monograph Wildgen (2023). Passages in this contribution correspond literally or in content to passages in sections 2.2 and 2.3, p. 17–22 of Wildgen (2023). This general advice must be sufficient, as explicitly marking all parallels and differences would have impeded the reading.

1 One must distinguish the very fast and automatic reaction of the sense organs to the stimulus and the elaborated reactions in the sense-related cortical areas (for instance V1 in the visual cortex). In the following this elaborated stage is understood as "perception."

2 Cf. Wildgen (2023, 197–200) for a comparison of the notion "symbolic form" in Cassirer and the notion of "sign" (symbol) in Peirce. "Cassirer's philosophy aims at a philosophy of human culture that addresses symbolic forms such as language, myth, science (later art, ethics, and technology). The meaning of the symbolic forms is given by the internal "reference structures" of consciousness; i.e., mental entities of different provenience and character are brought together. This trend is fully developed in human cultures and concentrated in a plurality of symbolic forms. The analysis of this plurality and the interrelations between the basic types is at the heart of Cassirer's theory of culture."



odors. Even if these labels are individually grounded in olfactory experience, olfactory communication remains vague if no social coordination has been achieved. The grounding of odors is difficult because the chemical structure of odors is very complex, and behavioral reactions are very context-dependent. Nothing corresponds to the Munsell classification of colors in the domain of odors (cf. Dubois, 2021: 203f).

Condillac (1754/1962) mentioned the problem of organization of the field of odors and the fact that it depends on the conditions and contexts of memorization.<sup>4</sup> Thus, if one experiences a series of odors, e.g., of flowers, coffee, and a nearby factory, a network that links and partially mixes these impressions and their evaluative reactions is created in memory. As a result, the neighborhood of odors in time and space, the amount of attention paid to them, and the pleasure or displeasure associated with them change from one individual to the next.

To summarize, the inventory of linguistic devices to account for 'odors' as a sensory experience, when extended beyond the search for 'basic terms', reveals that olfaction is not a 'mute' sense. The lack of lexical items in olfaction depends on sociocultural and linguistic constraints of 'talking about odors' rather than physiological constraints. A large diversity of morpho-syntactic, syntactic, and discursive devices is at work, differing from the (predominant lexical) categories relevant to vision (Dubois, 2021: 229).

### 1.3 The mode of communication referring to colors

Condillac 1754/1970 argues that the perception of colors alone does not constitute colored places, situations, motions, or objects. Instead, our visual brain allows us to see two or three colors and journey from one color to another. Eventually, the perception of colors contributes to an idea of extension, a space of colors.<sup>6</sup>

The astonishing differentiation of human color vision does not automatically entail a rich system of "ideas" of colored surfaces in our minds. Instead, attention, memory, the color space (surface) organization, and paths of visual attention are necessary to produce a selected set of "color ideas". Higher levels are only achieved if other types of sensibility are integrated and form an organization linked to objects, events, and actions. As the divergence of color terminologies in the world's languages shows, the need for a basic (stable) set of color terms varies between languages/ethnic groups, even on a similar level of cultural development. Therefore, communication about colors is not a universal and central concern of linguistic communities. However, this restriction does not preclude that the cognitive relevance of color perception is very high in individual perception, i.e., perception and communication do not share the same functional

pattern and do not respond to identical (or similar) needs. Instead, they follow their specific principles and laws.

In anthropological linguistics, Berlin and Kay (1969) have examined the elementary color vocabularies of different language cultures and the color values assigned to color words. A comparison with the physiological data resulted in a differentiation hierarchy. Color physiology can foresee the options if the degree of differentiation of color terminology increases in one language or in the case of a transition from one language to another.<sup>7</sup> One can infer that, to a certain extent, culture-invariant color vision determines the basic color lexicon. This result speaks for a weak cognitive determinism.

Although the universality of color terms or their sequence of replenishment remains controversial, it became clear that putting the perception of colors (in context) in words or utterances is a complex activity that is, in many cases, successful. Therefore, skepticism regarding the linguistic realization of primary color perception is not warranted. Instead, color perception and the perception of spatial contours, lines, and other visual features can achieve the status of symbolic forms in a semiosis that uses more general resources.

## 2 Multimodal integration and symbolic forms

In a given linguistic community, some sub-communities, i.e., painters or perfumes professionals, develop specific teaching competencies and techniques or even argue about odors/perfumes or colors. This practice modifies the communication demands. Our reflections point to three significant problems or problematic transitions between "sensibility" (perception) and "sense" (meaning) in a symbolic form, e.g., language):

- The transition from single sensibilities, e.g., olfaction or color sensibility, to an integrated perception is crucial.
- The transition between presentation and representation must consider the effect of attention, memory, and spontaneous imagination.
- The transition between perception, governed by principles of human neural architecture, and the dynamics of human communication, based on social interaction.

The most dramatic transition and the evolutionarily most recent one is that of (b) to (c). At level (c), specific human capacities are concentrated in the plurality of "symbolic forms" discussed by Ernst Cassirer. The problem of multimodality has been noted and discussed in philosophy at least since the 18<sup>th</sup> century; see our remark on Condillac in section 1. In the twentieth century, the philosopher and historian of science Ernst Cassirer developed a philosophy of symbolic forms and human culture. His notion of the symbolic forms is a good background for theories of multimodality insofar as the relative autarky of different modes and their interaction is put to the fore. Although the symbolic forms of language and theoretical/mathematical form-giving in the sciences are considered exemplary,

6 In Petitot (2017: 266) the neural integration of retinal opponent cell operations and positional information, mainly spatial frontiers is discussed in detail. "So from the lowest levels of the V 1 and the V 2 areas, there is a functional entanglement between the spatiality of perceived scenes and the colours of objects making them up. The brain must reconstruct by inference from the colour the objective reflectance of the surfaces perceived and this independently of a host of extremely variable factors such as the illumination of sources, indirect irradiation, angles coincidence, and reflection."

7 The color project was continued by Kay and others. Cf. for a later stage: Kay and Maffi (2013).

the other symbolic forms, e.g., art, myth, technology, ethics, and further ones, exist in parallel and may be older and thus the evolutionary sources of language and science. Therefore, we shall discuss the role of language on behalf of the more basic forms like odor- or color communication with specific reference to Cassirer's notion of "symbolic form."

## 2.1 What are symbolic forms?

Cassirer introduced "symbolic form" in his "Philosophy of Symbolic Forms," published in three volumes between 1923 and 1929. It contains two constituents: "symbolic" and "form." The second term includes the notions of *morphè* (Greek: shape/form) and its dynamic counterpart *morphogenesis*. The first term, symbolic, refers to symbol, a polysemic notion used in philosophy and aesthetics. A discussion of the range of meanings of this notion would ask for an independent treatise. However, an appropriate starting point is the notion of symbol in the theory of signs proposed by Charles Sanders Peirce in the second half of the nineteenth century. Cf. for comparing the contributions of Peirce and Cassirer, Wildgen (2023, Chap. 7).

Peirce considers three fundamental aspects of the sign, the central notion of semiotics: icon, index, and symbol. In this constellation, the notion of the symbol has to be delimited concerning the two neighboring concepts of the index (existential reference) and icon (reference via similarity, conceptual neighborhood).<sup>8</sup>

This definition needs clarification. The "dynamic object" is, in Peirce's terminology, the real-world object, the ultimate intention of the sign in its real-world usage. Peirce's statement tells us that the symbol only indirectly relates to the real-world object ("in the sense it is interpreted"). The interpretation depends on dispositions, habits, or conventions (see fn. 5). These determinations introduce a moment of arbitrariness; they depend on chance or, eventually, on many minimal causes beyond rational control. Regarding dispositions and habits mentioned by Peirce, one can assume rules of behavior that were acquired but gained law-like significance. In the case of conventions, these forces include cooperative effects in a community and conformity in social behavior.

## 2.2 Complex symbolic forms: the example of music and visual art

Immediate perceptual processes, such as those observable in olfaction and color vision, are still near to natural (bodily) morphogenesis. In contrast, colored surfaces and objects or artifacts (art) are symbols in the sense of the definition given by Peirce. In the case of odors, smell, and taste, very specific, institutionalized, and professional situations or contexts can lead to artificial norms and

devised terminologies that produce a symbolic level on which such perceptions may be efficiently communicated.

The symbolic forms of music and visual artifacts have a dominant founding in specific physical conditions: auditory perception and the motoric capacities to produce music in the first case and visual perception and motoric capabilities for creating visual artifacts in the second case. Language, an evolutionary late-comer, has a broad field of interacting capacities and needs many sources. The symbolic forms of music and visual forms can use the complexities of language to elaborate their repertoire and symbolic richness culturally. The multimodality of visual, acoustic, and linguistic communication is a major concern in social semiotics based on the linguistics of Halliday and further developed by van Leeuwen, Kress, Bateman, Wildfeuer, Hiippala, and others.<sup>9</sup>

The list of symbolic forms or communication modalities may be subdivided, although the evolutionary continuity implies an underlying continuum. As we showed in section one, odor/smell and basic color distinctions are at the threshold of symbolic forms. Visual artifacts and musical performances are still firmly rooted in perceptual patterns and motoric routines, i.e., bodily controlled (embodied). In contrast, languages are rooted in different perceptual domains, emerge or co-evolve from musical behavior (cf. Wildgen, 2018: 62–78), and refer to visually rooted (virtual) spaces. These levels range from the perceptually dominated odor/smell, still bodily rooted visual and musical forms to the more abstract and, to a large degree, culturally transmitted and sophisticated linguistic forms. Beyond, we find secondary symbolic forms heavily dependent on the three layers mentioned above but with new functions and more specific effects, for instance, myth and religion (cf. Wildgen, 2021) and the symbolic forms, technology, and ethics that Cassirer has added to his list of symbolic forms (cf. Sandkühler et al., 2003: 34f, 42f, and chapter 9). In the analysis of multimodality, it seems primary to consider the visual, musical, and linguistic symbolic forms. Multimodal artifacts or performances may combine or integrate two or three of these forms:

- The couple: visual and musical forms. Simple analogies concern the visual ornament in paintings or architecture and the ornamental enrichment of a melody. Cf. for examples and analyses Wildgen (2018: 97–102). Some painters like Paul Klee (1879–1940) have reflected the analogies between music and graphic or color design (cf. *ibidem*: 102–106).
- The couple: visual and linguistic forms. The close relationship is not only demonstrated by the evolution of writing based on pictorial representations but also on parallel vocations, i.e., painters who are poets and poets who are painters. Principles of symmetry and spatial ordering are valid for visual artifacts and poetry (cf. Wildgen, 2013: chapter 10). Stories and literary fiction can be realized via language (spoken or written), illustrations or sequences of pictures (cf. comic strips), and movies.
- The couple: musical and linguistic forms. The simplest integrated form is given with songs combining melody and lyrics. Poetry, with its rhymes and alliterations, the "concert" of vowels and consonants, and rhythmic features rival music or interfere with it.

<sup>8</sup> In a letter to Lady Welby on October 12, 1904, Peirce summarizes his position (developed after 1867, as he tells her; cf. Wiener (1958: 391). "I define a Symbol as a sign which is determined by its dynamic object only in the sense that it will be so interpreted. It thus depends either upon a convention, a habit, or a natural disposition of its interpretant (that of which the interpretant is a determination)." (*ibid.* 291f).

<sup>9</sup> For an overview of web design applications, comics, film, audio-visual materials, and video games, cf. Bateman et al. (2017) and an introduction to social semiotics Kress and van Leeuwen (1996).

- The triple: music, visual art, and language is typical for performances of musicians on stage, opera, and musicals and, since the 19<sup>th</sup> century, for movies. This topic will be the main concern in the section 3.

## 2.3 Language depends on the self-organization of perceptual capacities and their multimodal integration

Self-organization is a principle formulated in the cybernetics framework (Ashby, 1947) and involves searching for a stable state in a deterministic system. As already programmatically expressed by Wiener (1951), it is extrapolated from physical to biological, eventually symbolic systems. The purely syntactic problem of chaining elements of an existent vocabulary does not require a specific endowment and evolutionary processes enabling it; it can exploit much older sequencing techniques in motion and action. The real problem is semantic compositionality because the composition or blending of spaces with different topologies and the account of verb dynamics is crucial for sentential units. This tremendous problem must be resolved to allow stable and reliable communication via phrases and sentences. To arrive at a conventionalized system of linguistic behavior, early humans had to consider two major factors:

- a The cognitive demands for a stable solution of semantic compositionality,
- b The communicative and social demands for a compositional level of referentiality.

The solution to this problem is the gain of the evolutionary game called human language. Human utterances are, however, not restricted to isolated sentences. On the contrary, natural units are sequences of sentences, turns in conversation, adjacent pairs as in question–answer, and narratives or arguments. Therefore, human evolution has created the human language for its effective use in social communication, not for correctly using sentences or words.<sup>10</sup>

## 3 Multimodality in movies

Movies are an intriguingly complex example of multimodality, historically only comparable to theater and opera. In the following, we recapitulate major results obtained in analyzing movies in Wildgen

(2015, 2017, in English, 2013: chapter 7, and, 2018: 107–112; in German).

### 3.1 Space, movement, and narrative in movies

Movies are, on the one hand, moving images and visual communications in time. On the other hand, a story is being told. The balance between visual attraction and narratively motivated action differs in film genres. In action films, the focus is on the dramatic action focused on just a few actors, but it must be “woven” into a narrative texture. The public perceives many action films as part of a film series (see the series of James Bond movies). The narrative thread must, therefore, point beyond the respective film, provide structural analogies to previous films, and possibly prepare the next film.

### 3.2 The visual mode of movement in space

The location of the plot is the anchor for what is happening in the movie and makes it appear believable. In addition, characters and actions only become understandable and effective as constructs in the context of these places. In a broader sense, locations also include the costumes and location-specific behavior of the characters. In this respect, the film’s basic structure is already established with the exact construction of the locations and the directing of the courses of events and actions at these locations. Places of transition also play a major role, such as hotel lobbies, elevators, train stations, airports, and crowded squares (cf. the analysis of James Bond movies in Wildgen (2015, 2017).

The frame, the viewing window, plays a decisive role. A film in the Academy format (square) emphasizes the center more, thus increasing the illusion of depth. The broadband format emphasizes the horizontal, giving the landscape and storyline greater prominence. Actions and movements across the camera window can be followed longer without changing the setting. For example, suppose the film is set in architectural interiors. In that case, elements of architecture: doors, staircases, windows, narrow corridors, room dividers, and even furniture can create specific frames within the format and thus shape the space structure. People can be assigned to individual room segments. These spatial divisions can be repeated when moving through a suite of rooms. The person’s (and the camera’s) line of sight can be downwards (from a balcony, an upper floor window into the yard, onto the street) or up into a stairwell or, particularly extreme, into a rock face when climbing (correspondingly down into the dizzying abyss). In connection with the division of space, people and their actions acquire specific meanings. Structuring the space (particularly through the dividing lines and thresholds) is meaningful since it creates spatially separated binding structures of thematically related subfields (cf. Saint-Martin, 1990: 208ff). In the film, the spatial structures are transformed by the movement of the people and the camera. The film can even be viewed as a medium of spatial transformation. The moving person can be focused in the foreground; the surrounding space flows past the person. This feature is particularly evident in older Hollywood films and in some films of the New Wave, where the actors are filmed at the wheel of a car over long takes. However, the movement can also be caused by the camera

<sup>10</sup> The correctness criterion, traditional for school grammar, is linked to social conformity (obedience to rules) and sharply delimited social identities. Chomsky’s notions of grammaticality and the construction of “competence in a native speaker” that enables judgments of grammaticality is an abstraction that neglects the content of utterances in favor of formal features of concatenation (syntax). The intuitive notion of acceptability used by Chomsky’s teacher, Zellig Harris, was much easier to measure empirically and was nearer to the traditional notion of correctness. Indirectly, one could argue that Chomsky’s notion of competence is akin to traditional (prescriptive) school grammar. Cf. for current research on the notion of competence Vulchanov et al. (2022).

moving or the setting being varied from a long shot to a close-up. The cameraman plays a significant part in constructing meaning. The sequence of scenes and actions in different spatial segments is performed in the montage (in the editing room). Camera settings and the construction of the sequence of scenes in the editing room are thus the main organizational level of cinematic meanings. We can thus distinguish three sub-levels of the meaning construction of characters in the film:

- 1 The construction of meaning in the scene in front of the camera (prepared in the script, planned and controlled by the director, and specified by the actors).
- 2 The construction of meaning through the camera bears on the choice of setting, control of the lighting effects, and the camera's movement in space. In most cases, a multiple of the required film material is recorded, i.e., the camera creates a potential narrative space from which radical selections are made. Finally, complementary to the captured image is the off or hors-champ, which can be connotative.
- 3 The assembly of ready-made parts is either privative, i.e., large parts of the film material are discarded. The film director in the editing room corresponds to the sculptor who shapes the product that only exists in the imagination from a block of marble. Alternatively, the narrative order is created in the montage. In contemporary films, the scenes are reworked using computer technology and supplemented with special effects. In other cases, the main components of the entire film are generated electronically and supplemented by scenes recorded by a camera (the movements of real actors then serve as material for the animation of artificial characters, which are realized on the computer).

These three levels of meaning are essentially optical and visual. The textual-linguistic and the musical-acoustic dimensions can be added to them. Various versions can also be made with other texts in other languages or without integrated music, e.g., performed live by an orchestra. This autonomy of the different modalities makes the separability of the three basic levels, image – text – music, clear. As shown above, the visual level of organization is broken down into the organization in front of the camera (director's and actor's performance) – camera and lighting performance – editing, assembly, and special effects. The film must integrate these three levels and their sub-levels, i.e., put them together without too much redundancy and avoid disaccord or inconsistency. The integration occurs in special zones of the respective organizational level, so these remain relatively autonomous. The montage and the text organization must fit together. The modification by montage can change the narrative content and, thus, the textual level. The camera's focus must also be in harmony with a person's weight or role in the text. Suppose the main character is not emphasized visually by size or sharpness or in the movement. In that case, certain narrative threads are not adequately realized in which she/he is the center of attention. The integration of music must be coordinated with editing and montage. However, it is also tied to the narrative structure insofar as complication and climax passages of the narration correlate with the music. The dominant dimension is the visual construction, which the actors, camera, editing, and montage carry out. In addition to the continuous (imperceptible) cuts in Hollywood films, one can consider Eisenstein's formalistic montage,

Orson Wells' non-causal montage, and Godard's montage of the gap (cf. Agotai, 2007: 98).

### 3.3 Film music or the integration of music, (moving) images, and language

The music in the film can form a background without much effect and be devoid of any informational content.<sup>11</sup> The history of film, music, and literature shows:

- The film can be received without language (cf. the experience of silent movies).
- The music can also be satisfactory without speech, song, or accompanying text (cf. instrumental and electronic music).
- The text of classical literature can do without illustration by pictures (or even films) and musical accompaniment.

The central question will be how the visual medium on the one side and the narrative and descriptive dimension of the film on the other relate to the music's temporal-rhythmic and harmonic-melodic structure. The relationship can be a juxtaposition (an additive combination) or a selective interlocking. The rhythm of the scenes or cuts in the film and the phrases or melodic lines of the music can motivate a creative composition with emerging new qualities. The interaction can be parallel or contrary (contrapuntal). Within the three-fold relationship, music–film–language, two-way relationships can also appear. Music and language can be represented by songs, which are then embedded in the film or even form the main theme. Filmed dialogues integrate image and language and can be embedded in a musical context. The film can also have a musical or an opera as the subject and show a combination of music, language (e.g., in songs), and visually presented actions. Finally, the film can refer to other films or film-making, and the music of another film can be quoted.<sup>12</sup> Different combinations or blends can be conceived:

- 1 *The juxtaposition of music and film*: in the silent film era, the effect of the darkened rooms, the ghostly light displays, and the projectors' noise could be masked and mitigated by the music playing simultaneously (see Adorno and Eisler, 1944/1977: 116). The music used was mostly classical-romantic piano and salon music (cf. Kreuzer, 2001: 26). Both in the visual and in the musical area, a conventionalization of the means of expression leads to the cliché, i.e., the means are known from other contexts and uses and thus lose their current

11 A chapter in Wildgen (2018: 107–112; in German) treats the role and development of music in movies. The monograph discusses general and detailed aspects of the interaction of music with language and art.

12 Piel et al. (2008: 65f) refer to the film: First Name Carmen (original title: Prénom Carmen) by Jean-Luc Godard (in 1988). The protagonists pretend to be making a movie (but are planning a bank robbery). In the film, there is an orchestra whose violist is, in turn, a character in the film. Beethoven's string quartet, which is being rehearsed by the orchestra, also provides the film's structure. Even more often, such a cross-reference between music and film is witnessed in films about musicians.



expressiveness. From this perspective, Adorno and Eisler (1994/2007) criticize the film industry (especially in Hollywood before 1930). The reuse of motifs from the classical music of the 19th century does not correspond to the historical context of the film viewer, as it reflects intellectual and aesthetic movements in the 19th century. The firmly established classical musical forms used in different cinematic and narrative contexts are redundant or meaningless in these contexts.<sup>13</sup>

- 2 *The partial integration of film and music*: it resulted from the desire to combine scenic/visual and musical expression. From 1909, technical and commercially viable proposals for solving the problem came into use. Types of scenes in the film were associated with musical examples with corresponding dynamics or an appropriate pace. In the successful era of Hollywood films (1908–1927), “cue sheets” appeared, i.e., a script that assigned music scores to the film’s scenes. The composers began composing their music adapted to the individual film.<sup>14</sup> The elements of the musical tradition that were processed for the film mostly came from Richard Wagner, Richard Strauss, Giacomo Puccini, Giuseppe Verdi, Gustav Mahler, Claude Debussy, Maurice Ravel, and Alexander Scriabin. The technique of integrating music and cinematic events first resorted to the technique of leitmotifs that Richard Wagner had developed for the opera.<sup>15</sup> Recurring motifs or melodies are assigned to important characters or locations, thus reinforcing the movie’s cinematic and textual coherence.<sup>16</sup> Max Steiner introduced the accompanying music to the dialogues. Specific solutions had to be found for the technical problem of an exact adjustment of the film and musical sequences, e.g., the use of stop signals for the orchestra and the conductor of the film music or even the writing of the film music according to a stopwatch. When both the film and the soundtrack could be cut and put together in montages, and especially since the components of the film were ordered on the computer or since digital production, there was no technical obstacle to the temporal coordination of film and music (through so-called temp tracks).
- 3 *The complex integration of music and film*: instead of supplementing the sequence of images with a sequence of musical sequences, the music can also be organized in

opposition to the image, as in Hitchcock’s film *Rebecca* (USA 1941), for which Franz Waxman wrote the music. The character, e.g., *Rebecca*, may be absent, yet the musical motif represents her. The leitmotif linked to a person or a location can be replaced by a theme that runs through the entire film and gives it unity; compare the film: “*Play Me the Song of Death*” (Italy/ USA 1968, director: Sergio Leone). The film music was composed by Ennio Morricone (1928–2020). The title melody on the harmonica is embedded in the sound of a symphony orchestra with a choir. Morricone thus created a new standard for film music that was valid until the digital film era.

- 4 *Specific musical colors and moods*: with the focus on psychological aspects in film, moods were increasingly reflected through music; even a “mood technique” was developed. In the context of “Film Noir,” dissonances played a role in controlling emotions.<sup>17</sup> Miklós Rózsa (1907–1995) portrayed the psychological violence of criminal characters with the help of dissonant music, for example, in “*Double Indemnity*” (USA, 1944; director: Billy Wilder). The film music thus became more and more independent of the visually presented events. For example, the music for the film “*Star Wars*” (USA 1977; directed by George Lucas, music by John Williams) was largely realized as orchestral music. It claimed to present the narrative pattern parallel to but independently of the film through synthetically modulated noises. An economically desirable consequence was that the music could be marketed independently of the film.
- 5 *Musical innovation in film music*. Hitchcock worked closely with his composer Bernard Herrmann (1911–1975) during the production process. The subject of vertigo is represented both in the filmed action and in the music (cf. Kalinek, 2007). The central motif consists of arpeggios played in opposite directions by the discant and bass voices without creating a stable melody. The beginning and end of the motif are characterized by sevenths and seconds, i.e., by dissonant intervals. This pattern is later rhythmically doubled, i.e., accelerated and marked by great changes in dynamics from *ff* (*forte forte*) and *pp*. (*piano piano*). In addition, the tempo moves from *accelerando* to *ritardando*. The musical swirl’s intensity is increased by shifting the focus from the first to the second beat. In one scene, the swirl (or vortex) is also rendered visually by a rotating-colored spiral, i.e., the musical structure receives a geometrical and visual equivalent.

Film music is historically dependent on the overall development of music despite its autonomy in the phase of silent movies. Under the influence of digital technologies and the emergence of easily accessible libraries of musical motifs or passages on the internet, this has become a general characteristic of modern music culture (cf. Vernalis, 2013). On the one hand, recycling classical-romantic music traditions

13 The use of musical clichés is the consequence of industrialization in film production and the commercial pressure on film composers, who had to forego the risks of artistic innovation for the sake of box office revenue. Film music, therefore, lagged decades behind the modern development of music.

14 Important composers and arrangers in Hollywood were: Max Steiner (1888–1971), Erich Wolfgang Korngold (1897–1957), and Alfred Newman (1900–1970).

15 The leitmotifs are also taken from other films. Max Steiner gave his orchestrator brief instructions as to which motifs from which films he should use (cf. Wegele, 2010: 20). In contrast to the early phase of the film, Hans Zimmer writes musical suites for the planned film even before shooting begins. They are then adapted to the film at the end using modern methods or the film is adapted to the music.

16 See Weill (1946/1990: 135). Film music was shaped by this successful style of the early days well into the 20th century. However, electronic music gradually dissolved this basic pattern (see the work of Klaus Zimmer, born in 1957).

17 Ennio Morricone mentions in an interview that filmgoers will likely accept dissonant music in brutal and shrill horror films. His attempts to use modern music as film music usually failed due to the directors’ lack of understanding. They pointed out to him that films tend to be produced for ordinary people.

reinforced the public consciousness of classical music. On the other hand, a compilation of the most diverse musical styles became possible. Art music is mixed with pop music, and synthetic auditory sounds are combined with speech sounds.<sup>18</sup>

The fact that the semiotic structures of the music, like those of the (moving) image, are self-sufficient makes their connection, their semiotic “blend,” particularly attractive because the transitions and the mutual complementarity allow the emergence of new meanings in the interaction of the two symbolic forms. The new media of television, video, and the Internet also benefit from this trend. However, a longer historical phase was necessary before the potential of an effective combination of film (image) and music (sound) finally led to satisfactory results and innovative developments.

## 4 Conclusion: multimodality and semiotics

The rivalry of arts, mainly poetry, painting, and music, has been a topic since antiquity. In modernity, Gotthold Lessing (1729–1781) focused on the rivalry between poetry and painting in his essay *Laocöon. An Essay on the Limits of Painting and Poetry* (1766). The underlying question of the contribution of our senses: touch, odor, sight, and hearing to human understanding had already been thoroughly treated by Etienne de Condillac (1714–1780) in two books (*cf.* section 1). Both founding fathers of semiotics (*sémiologie* in Saussure’s terms), Ferdinand de Saussure and Charles Sanders Peirce, tried to generalize their original field of study (before and after 1900). Saussure looked beyond language, and Peirce looked beyond philosophical logic. In this move, other types of semiosis came to the fore. However, they remained secondary to language (Saussure) or in comparison to logic (as a formal language) and science (Peirce). For the founders of semiotics, the question of multimodality remained secondary. It was the philosopher Ernst Cassirer (1874–1945) who addressed the interaction and conflict of different forms of semiosis (called “symbolic forms”) in his trilogy on language (vol. 1), myth (vol. 2), and science (vol. 3) in the twenties of the last century. Later, he enlarged the field to cover art, technology, and others. A major concern of this contribution is the suggestion that the modern discipline of multimodality studies may find a proper framework in the philosophical traditions from Condillac to Cassirer.

In the main sections, the interaction, rivalry, and conflict between language (text), art (painting), and music were the main topics. At the heart of it lies the question of “meaning” or “signification” in the three modalities.<sup>19</sup> Major questions are:

- How can “meaning” and its semiotic realization be compared across modalities? Is there a common denominator?

- How is the apparent complexity of human products in these fields organized? Are the architectures of rich “meaning compounds” comparable? How far is an integration or blend of the different modalities possible?

The answers in traditional structuralism since Bloomfield, Hjelmslev, Chomsky, and others either avoided the question of (referential) meaning because it seemed to be ontologically freighted or relegated it to some intuitive logic (Hjelmslev, 1935) or formal ontologies in the case of intensional or possible world logics. Such strategies are of no avail in the treatment of visual art or music and only a meager substitute in the case of language. The problem of “meaning” remains the heart piece of multimodality semiotics.

The surface forms (in morphology and syntax) may be reduced to simple base forms and rules of concatenation and transformation (*cf.* Chomsky, 1957). Hockett (1954) had already distinguished two techniques of linguistic analysis: Item and Arrangement (I.A.) and Item and Process (I.P.). Both can be exemplified for languages and can easily be applied to musical and visual forms, for instance, temporal sequences of musical phrases and planar or spatial organization of elements in paintings and architecture. The different modalities differ mainly in the dimensionality of the organization in space and time. The semantics of signs or sign complexes cannot be reduced to such simple principles. They establish not only a link to the complexity of nature (the world surrounding humans) but also to cognitive/mental spaces, emotional reactions (feelings), and functional practices giving sense to symbolic behavior. In these respects, the modalities are very different. Although human language has, in many respects, a denotative function embedded in emotional/cognitive and practical contexts, music can only, under specific conditions, have denotative functions, for instance, in program music. Visual artifacts (art, architecture, film) may be illustrative and thus come near to the denotative function of language. However, as demonstrated in abstract paintings or instrumental music, they are not fundamentally denotative and can be ripped off this function.

The real challenge of multimodality lies in the difficulty of blending, integrating, or comparing the meaning effects of different modalities. The spaces of meaning have different organization and content; one cannot just arrange these contents in the same manner as the elements in a linear sequence of phonemes or morphemes.<sup>20</sup> A further difficulty consists of the dependence of multimodal semiosis from contexts that are either shared in a culture or fixed in use situations. This aspect was highlighted in social semiotics referring to multimodality by Bateman et al. (2017).

## Author contributions

WW: Writing – original draft.

<sup>18</sup> A special genre of music has emerged in the context of computer games. The music takes on new functions. The American musician Garry Schyman, who composed the soundtracks for the *Bioshock* series, writes: “Our compositions accentuate and expand the story. The moods we create are part of the gaming experience.” Graff (2017: 9, col. 4).

<sup>19</sup> The topic of “meaning” in art, music, myth and language has been treated in several publications by the author, recently in Wildgen (2023).

<sup>20</sup> *Cf.* Brandt (2004). The blending of semantic domains is only an extended device using logical operations of selective union regarding sets of features. A more realistic model should respect the topology of different domains and check the coherence of mappings between local spaces. The proper background is given by dynamic systems theory and not logics.

## Funding

The author declares that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

- Adorno, T. W., and Eisler, H. (1944/1977). "Komposition für den Film" in *Hanns Eisler. Gesammelte Werke* (edited by Stephanie Eisler and Manfred Grabs) (Berlin: Henschel).
- Adorno, T. W., and Eisler, H. (1994/2007). *Prejudices and Bad Habits*, reprinted in Dickinson, K. (ed.) (2007) *Movie Music. The Film Reader*. London: Routledge, 25–35.
- Agotai, D. (2007). *Architekturen in Zelluloid. Der filmische Blick auf den Raum*. Bielefeld: Transcript.
- Ashby, W. R. (1947). Principles of the self-organizing dynamic system. *J. Gen. Psychol.* 37, 125–128. doi: 10.1080/00221309.1947.9918144
- Bateman, J. A., Wildfeuer, J., and Hiippala, T. (2017). *Multimodality foundations, research, and analysis: A problem-oriented introduction*. Berlin: De Gruyter Mouton.
- Berlin, B., and Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley: Berkeley U.P.
- Brandt, P. A. (2004). *Spaces, domains, and meanings*. Bern: Peter Lang.
- Cassirer, E. (1995). *Philosophie der symbolischen Formen*. 3 vol. [first edition 1923/1925/1929]. Darmstadt: Wissenschaftliche Buchgesellschaft [English new translation: The philosophy of symbolic forms. London: Routledge, 2020].
- Chomsky, N. (1957). *Syntactic structures*. Den Haag: Mouton.
- Condillac, E. (1746/1982). *Essai Sur L'Origine des Connaissances Humaines: Ouvrage où l'on réduit à un seul Principe tout ce qui concerne l'entendement humain*. Paris: Pierre Mortier Translation by Franklin Philip, essay on the origin of human knowledge, in: *Philosophical writings of Etienne Bonnot, Abbé de Condillac*, Volume II, Hillsdale NJ: Lawrence Erlbaum, 1982.
- Condillac, E. (1754/1970). *Traité des sensations*. London/Paris: Bure. Slatkine reprints <https://gallica.bnf.fr/ark:/12148/bpt6k96333766/f12.item.texteImage>, 1970.
- Dubois, D. (2021). *Sensory experiences: Exploring meaning and the senses*. Amsterdam: Benjamins.
- Graff, B. *Pokémon Symphony. Der Sound von Computerspielen wird so sorgfältig komponiert wie Filmmusik*. Mit einem Unterschied", *Süddeutsche Zeitung*, 98 (28th of April, 2017), part 2: 9, rows 1–5. Munich: Süddeutscher Verlag, 2017.
- Hjelmslev, L. (1935). *La catégorie des cas. Étude de grammaire générale*. New Edn. München: Fink.
- Hockett, C. (1954). Two models of grammatical description. *Word* 10, 210–234. doi: 10.1080/00437956.1954.11659524
- Kalinec, K. (2007). "The language of music" in *A brief analysis of Vertigo*. ed. K. Dickinson (London: Routledge), 15–23.
- Kay, P., and Maffi, L. (2013). "Number of basic colour categories" in *The world atlas of language structures*. eds. M. Dryer and M. Haspelmath (Leipzig: Max Planck Institute for Evolutionary Anthropology). Available at: <https://wals.info/chapter/133>
- Kress, G. (2010). *Multimodality: a social Semiotic approach to contemporary communication*. London: Routledge.
- Kress, G. R., and van Leeuwen, T. (1996). *Reading images: The grammar of visual design*. New York: Routledge.
- Kreuzer, A. C. (2001). *Filmmusik. Geschichte und Analyse*. Frankfurt/Main: Peter Lang.
- Locke, J. (1690/1975). *An essay concerning human understanding*. Oxford: Clarendon Press, First publication, Eliz Holt, London 1690; cf. Available at: <https://www.gutenberg.org/files/10615/10615-h/10615-h.htm>.
- Petitot, J. (2017). *Elements of Neurogeometry*. Functional Architectures of Vision: Springer, Cham.
- Piel, V., Holtsträter, K., and Huck, O. (Eds.) (2008). *Filmmusik. Beiträge zu ihrer Theorie und Vermittlung*. Hildesheim: Olms.
- Saint-Martin, F. (1990). *Semiotics of visual language*. Indiana: Indiana University Press.
- Sandkühler, H. J., Pätzold, D., Freudenberger, S., van Heusden, B., Plümacher, M., and Wildgen, W. (2003). *Kultur und Symbol*. Ein Handbuch zur Philosophie Ernst Cassirers, Stuttgart: Metzler.
- Vernalis, C. (2013). *YouTube, music video, and the new digital cinema*. New York: Oxford U.P.
- Vulchanov, V., Sorace, A., Suarez-Gomez, C., Guizarro-Fuentes, P., and Vulchanova, M. (2022). The notion of the native speaker put to the test: recent research advances. *Front. Psychol.* 13:875740. doi: 10.3389/fpsyg.2022.875740
- Wegele, P. (2010). Max Steiner und die Filmmusik des Golden Age in Hollywood: Eine kurze Betrachtung der wichtigsten stilistischen Merkmale anhand der Musik Steiners zum Film. *Kieler Beiträge zur Filmmusikforschung* 6, 8–36. doi: 10.59056/kbzf.2010.6.p8-36
- Weill, K. (1946/1990). "'Music in the movies', Harper's bazaar" in *Musik und Theater. Gesammelte Schriften. Mit einer Auswahl von Gesprächen und Interviews* (New York: Hearst).
- Wiener, N. (1951). *Cybernetics: or control and communication in the animal and the machine*. Paris: Hermann & Cie.
- Wiener, P. P. (Ed.) (1958). *Charles Sanders Peirce. Selected Writings*, New York: Dover Publications.
- Wildgen, W. (2013). *Visuelle Semiotik: Die Entfaltung des Sichtbaren: Vom Höhlenbild bis zur modernen Stadt*. Bielefeld: transcript-Verlag.
- Wildgen, Wolfgang, *Catastrophe theory and Semiophysics: with an application to movie physics, Language and semiotic studies* 1: 61–88 (Soochow University Press, China, (2015).
- Wildgen, W. (2017). "'Movie Physics' or dynamic patterns as the skeleton of movies" in *Film text analysis (Routledge advances in film studies)*. eds. J. Wildfeuer and J. A. Bateman (London: Routledge), 66–93.
- Wildgen, W. (2018). *Musiksemiotik: musikalische Zeichen, Kognition und Sprache*. Würzburg: Königshausen & Neumann.
- Wildgen, W. (2021). *Mythos und Religion: semiotik der Transzendenz*. Würzburg: Königshausen & Neumann.
- Wildgen, W. (2023). *Morphogenesis of symbolic forms: meaning in music, art, religion, and language*. Cham (C.H.): Springer Nature (Series: Lecture Notes in Morphogenesis),.



## OPEN ACCESS

## EDITED BY

Janina Wildfeuer,  
University of Groningen, Netherlands

## REVIEWED BY

Dusan Stamenkovic,  
Södertörn University, Sweden  
Joost Schilperoord,  
Tilburg University, Netherlands

## \*CORRESPONDENCE

Loli Kim

✉ loli.kim@ames.ox.ac.uk

RECEIVED 06 December 2023

ACCEPTED 22 May 2024

PUBLISHED 26 June 2024

## CITATION

Kim L and Calway N (2024) SFDRS as a metalanguage for 'foodscaping': adding a formal dimension to an interdisciplinary, multimodal approach to food.  
*Front. Commun.* 9:1351733.  
doi: 10.3389/fcomm.2024.1351733

## COPYRIGHT

© 2024 Kim and Calway. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# SFDRS as a metalanguage for 'foodscaping': adding a formal dimension to an interdisciplinary, multimodal approach to food

Loli Kim\* and Niamh Calway

Faculty of Asian and Middle Eastern Studies, University of Oxford, Oxford, United Kingdom

'Foodscaping' seeks to understand how meaning is made through humans' interaction with food in particular environments through a multimodal and interdisciplinary analytical lens. As part of a foodscaping project, researchers often interpret food environments to which they are not intimately 'local'. This presents cross-cultural limitations in the production of analysis. Most pertinently, how can personal interpretation be divorced from locally salient and meaningful discourses? This paper presents the findings of a pilot foodscaping analysis using the box notation style of Kim's Korean Segmented Film Discourse Representation Structures (K-SFDRS). K-SFDRS notation, developed to provide both coarser- and finer-grained formal transcription for South Korean multimodal film discourse analysis, is tested as an analytic tool for an authentic South Korean foodscaping experience. This paper aims to ascertain whether the formal nature of K-SFDRS transcription is a useful aid to the analysis of a foodscape, which otherwise risks relying heavily on personal interpretation. This pilot study presents an introduction to both foodscaping and (K-)SFDRS, outlines the potentials of (K-)SFDRS notation within a foodscaping context, offers a step-by-step outline for constructing K-SFDRS box notation using an exemplar South Korean foodscape, and finally demonstrates how this box notation may be used in the support of foodscaping analysis in various interdisciplinary channels. During this pilot study, the authors make a novel methodological development in the form of what they term 'cluster structures', which overcome the problems presented by the lack of cinematic narrative editing in spontaneous discourse, segmenting meaning into logical forms within which structures of meaning are hierarchised without requiring the discourse relations to structure the logical forms themselves in narrative discourse following the original K-SFDRS methodology. The paper concludes that K-SFDRS, alongside the aforementioned methodological development, has potential to help foodscaping researchers constrain interpretation to salient discourses and direct foodscaping analysis down meaningful avenues. Through its culinary scope, this chapter adds a new disciplinary dimension to discussions of metalanguage and makes an innovative contribution to the current corpus of multimodal research.

## KEYWORDS

foodscaping, Segmented Film Discourse Representation Structures, Korean Segmented Film Discourse Representation Structures, multimodal discourse analysis, formal notation



# 1 Introduction

In this paper, we present a pilot study whose aim is to test the hypothesis that Segmented Film Discourse Representation Structures (SFDRS), more specifically the Korean language and culture specific K-SFDRS, can be used to formally transcribe spontaneous and authentic foodscaping experiences. The paper will test whether this formal tool can productively be used to produce data in support of the analysis of a Korean foodscape and the multimodal threads that produce meaning therein. This paper tests whether foodscaping, a discipline borne from multimodal analysis, but which lacks a formal mode of data representation across modalities, benefits from the use of logical forms which incorporate a range of multimodal utterances.

## 1.1 Foodscaping

This paper is borne from the authors' work on a wider project concerned with 'foodscaping' various regions of East Asia, and our search for formal methodology in support of said project. In order to understand the basis of the current paper, we must first outline foodscaping and, accordingly, the gap in formal data analysis it presents. Norah MacKendrick defines a foodscape as such: 'Consider the places and spaces where you acquire food, prepare food, talk about food, or generally gather meaning from food. This is your foodscape' (MacKendrick, 2014, p. 16). A Foodscape describes a wider space centred around a food environment and in which people interact with food, not only by consuming it but also producing it, acquiring it, preparing it, and socialising around it. The crucial element in MacKendrick's above definition is the specification that in a foodscape, human actors 'generally gather some sort of meaning from food' (MacKendrick, 2014, p. 16). This consideration sits at the heart of foodscaping as the authors define it. Foodscaping is the study of foodscapes; more specifically, it is the methodological process through which a participant, or observer, analyses and unravels the multimodal strains through which meaning is derived from the foods in question in that specific space. Thus, foodscaping seeks to understand how meaning is made through humans' interaction with food – and, by extension, with one another over food – in particular environments (Calway et al., 2025).

Further to this, foodscaping puts primacy on the understanding of the cultures and societies which surround and define foodscapes. If meaning is to be derived from food and the manner with which it is interacted, a multitude of culturally- and contextually-informed values exert an important influence. For example, imagine two foodscapes based around establishments serving fried chicken: one in London, United Kingdom, and another in Seoul, South Korea. Whilst the basic building blocks of the foodscape may be similar, the meanings constructed around the foodscape are entirely different. The side dishes customers eat with the chicken, the times at which customers purchase and eat the food, the groups or individuals with which they choose to eat the food, whether the chicken is eaten in the restaurant or at home as a takeaway, the manner and language in which the food is ordered and talked about; these factors and more, whilst being unique to each individual, behave to a certain extent according to custom and cues unique to each location and culture. It is the interplay of these unique customs and cues which foodscaping seeks to depict.

In short, individuals and groups derive meaning from food. These meanings are unique to certain groups or individuals, interacting with certain foods, in certain places, at certain times. Beyond just describing the resulting food cultures, foodscaping seeks to lay bare the exact factors that have converged to produce said meanings. Foodscaping, on the one hand, seeks to separate and consider these factors on their own terms, whilst, on the other hand, simultaneously recognising that it is their confluence which ultimately results in the meaning-making inherent to food in society. Thus, through foodscaping, foodscapes are deconstructed into numerous multimodal threads, understood according to the relevant academic discipline (i.e., language is analysed linguistically, historical processes are understood historically, anthropological considerations are understood anthropologically, etc.), and ultimately reconstructed in order to produce a refreshed, full picture of the meaning of food within the foodscape. Foodscaping therefore avoids pitfalls of which current food studies often fall foul: considering food through the lens of only one modality or discipline; or, conversely, attempting to consider the resulting meanings of food without affording due focus to each of these multifarious factors.

This approach to the analysis of a foodscape is highly complex; multiple modal inferences must be considered (such as spoken language, written language, movement, smells, tastes, sights, sounds, and temperature), each of which can be understood according to various disciplines (history, linguistics, sociology, psychology, anthropology). Where, then, does one begin? And how does one separate the different, simultaneous processes through which meaning is derived? These are the key questions which began our investigation, the result of which is this paper. It is all very well to say that Koreans eating a barbecue bond with one another by sharing food, on the one hand, and retain social distinctions by pouring one another alcohol according to age, on the other, but how does one trace these two processes all at once (Yu, 2017). More importantly, how can one use data (rather than just culturally-informed intuition, which is susceptible to bias) to verify that these processes are indeed happening? And lastly, how do we use this data to understand which threads of analysis are the most salient to the meaning-making we observe? Foodscaping, and food studies at large, lacks a logical approach through which foodscapes can be formally analysed and through which these questions can be answered using verifiable data. This is where KSFDRS comes in.

## 1.2 (K-)SFDRS

Before explaining the potential applicability of K-SFDRS to Foodscaping, we must first outline K-SFDRS, as well as the SFDRS from which it was derived. 'Segmented Film Discourse Representation Structures' (hereafter SFDRS) are a formal means of transcribing multimodality and how it unfolds to construct discourse in film, developed as a part of a framework from Multimodal Film Discourse Analysis by Wildfeuer (2012, 2014). The layered, dynamic discourses in foodscapes share a parallel with SFDRS in this respect, and have encouraged piloting the framework in this paper.

K-SFDRS is distinct from SFDRS in that it uses a set of Korean-specific socio-pragmatic rules for verbal and non-verbal language ('socio-pragmatic primitives') developed from Kiaer and Kim (2021) in addition to audio and visual elements to identify salient modalities and to infer the defeasible eventuality of those modalities as they

interact, using them to draw the discourse structure. Studies by [Kiaer and Kim \(2021\)](#) and [Kim \(2022\)](#) both found that the socio-pragmatic expressions attached to modalities construct Korean narrative, drive it forward, and make its structure cohesive, as well as being predominantly responsible for defining the personalities, motivations, and intentions of characters. [Kim \(2022\)](#) found that discourse structures do not make sense without including socio-pragmatics in interpretation, and that doing so made interpretation much more specific. [Kiaer and Kim \(2021\)](#) have pointed out failing to follow socio-pragmatic rules, a lack of corroboration among modalities, or inconsistency in the socio-pragmatics at play in a given orchestration of expressions, either does not make sense or is purposely employed to show insincerity or strangeness. K-SFDRS also employs [Kim and Kiaer's \(2021\)](#) granular division of discourse, which they term 'higher activities,' to organise the parts that form discourse and finer and coarser levels, though we do not employ this division system in this preliminary investigation. In order to illustrate the socio-pragmatic elements that constrain Korean interpretation of modality, which must be considered when studying Korean modalities, K-SFDRS includes in its annotation '[k]' in addition to marking referents as either audio '[a]' or visual '[v]'. This '[k]' is positioned beneath people, objects, and expressions to indicate that the above is a Korean-specific socio-pragmatic or discursive element that has implications on other modalities in the given meaning-making process. In other words, '[k]' indicates that an element is meaningful in the given context because of Korean socio-pragmatics. The use of the '[k]' notation is exemplified in section 2 of this paper.

### 1.3 Why apply (K-)SFDRS to (Korean) foodscaping

(K-)SFDRS is highly applicable to foodscaping research in two key ways: firstly, multimodality is laid bare, and secondly, the most salient instances of multimodal communication can be identified in a controlled manner. Furthermore, it enables both researchers and readers to verify the findings of the ultimate output, referencing the specific modalities in their original context against the final interpretation of the author.

#### 1.3.1 Multimodality

(K-)SFDRS provides a means of formally transcribing multimodality to facilitate their full and proper analysis. (K-)SFDRS notation not only takes into account all modalities present in a section of discourse, but it converts them into logical forms which point to the mode in which it is manifested, all whilst considering all modes together in one chronological graphic representation. In this way, (K-)SFDRS analysis avoids putting primacy on any one modality by considering them all as equal in their potential importance to meaning-making, whilst also retaining the nature of each modality so that they can be identified in the logical representation of the discourse. This limits the bias researchers might afford the more 'apparent' methodologies in a foodscape analysis, instead forcing them to consider all possible multimodal referents before narrowing them down to the most salient factors only after considering each and every one. We believe this to be highly relevant to foodscaping, as well as frameworks that resonate with it such as culinary linguistics ([Gerhardt et al., 2013](#)), where each modal thread must retain its original modality to be properly analysed, but should also be considered in terms of its

confluence with other multimodal utterances in working towards the formation of meaning.

#### 1.3.2 Salience

A major issue faced in discourse analysis is the difficulty distinguishing between salient and arbitrary ([Bateman and Wildfeuer, 2014](#)). Foodscaping is no exception. In fact, it is potentially one of the greatest challenges in discourse analysis, because it involves the interpretation of discourses in live environments, without the carefully planned narrative and editing to guide the researcher to meanings. Yet, these spaces exist at the communicative core of our social lives and are rich in customs, traditions, and forms of communication ([Kiaer et al., 2024](#)), and therefore undoubtedly contain discourses.

(K-)SFDRS has the basic aim of codifying and representing multimodal instances, enabling the researcher to identify the most salient moments in contributing to the segments of meaning inferred. We begin with the hypothesis that this may be a starting point for analysing spontaneous discourse in food environments. If people can identify salient modalities, then modalities that are acting in similar ways or corroborating with one another will also be identifiable; in the same way, it becomes apparent when modalities do not corroborate with one another. Based on this, referents may be categorised into groups which, although demonstrating subtle variations, demonstrate similar factors in the meanings to which they contribute. These groups may then be brought together into clusters that build more general meaning. As the researcher gains organisation over these clusters, they may employ their disciplinary background and individual expertise to draw discourses from the meaning-making processes. These structures can, with persistence, continue to branch further and further. The precise manner in which the branches develop depend on the discipline of the researcher; whilst the logical forms help to guide interpretation, the ability of the researcher is required to interpret said forms and thus their knowledge of these discourses is vital. This is why we recommend collaboration between researchers from different disciplines in research of foodscapes to make the most of the methodology (for more on this, see section 4).

#### 1.3.3 Logical forms/graphic representation

The production of logical forms to which the researcher may point, and the reader may consult, further makes (K-)SFDRS a potentially beneficial tool for foodscaping. The description of spontaneous discourse in a foodscape is difficult to achieve through written prose, particularly when the aim is to identify several discourses, which require the identification of several salient modalities. Furthermore, an understanding of how and when each event takes place is most likely germane to a reader's understanding. The box notations produced through (K-)SFDRS could prove useful in enabling the author to demonstrate this in concise, logical terms. Additionally, foodscaping, like a lot of ethnographic and cultural analysis, rests on the trust of both the researcher and the reader on the researcher's own interpretation of the target culture and society. Whilst the authors do not wish to cast aspersions on the credentials of ethnographic researchers (indeed, quite the opposite), we propose (K-)SFDRS as a useful tool for reasoning, reviewing, and verifying to the researcher and reader alike the legitimacy of the researcher's inferences. (K-)SFDRS effectively transforms qualitative interactions into a piece of data. Whilst this alone may not adequately give the full picture of a foodscape, it certainly can be used by the researcher as quantitative evidence for the veracity

of their observances: which modalities are most salient in this foodscape and which discourses do they take their salience from? Not only does the process of (K-)SFDRS notation help find the answers to these questions, but the graphical forms produced lay them bare.

### 1.3.4 Cross-cultural translation and interpretation

The key factor in our choice to apply K-SFDRS, rather than SFDRS, to foodscaping analyses in the Korean context is its ability to effectively ‘translate’ Korean meaning-making. Cross-cultural development of SFDRS through socio-pragmatics has already been demonstrated as beneficial to interpreting culture specific discourses; studies like Kim and Kiaer’s (2021) focused on the case of Korean filmic discourse. Bohnemeyer et al. (2007, p. 496) further proposes that ‘Given this intralanguage variability, we may expect a high amount of crosslinguistic variation in event representations’; K-SFDRS has the potential to approach this variability in foodscaping.

An analysis grounded in the original cultural context of a discourse is very important, especially for newcomers/non-natives (both conducting and reading) analysis of global foodscapes. Since SFDRS derives from SDRT (Asher and Lascarides, 2003), which was developed for the English language, there is much that does not translate and requires recognition of the Korean configuration of multimodality in order to make sense of discourses. This brings us to a matter of the utmost importance when analysing discourses of the Other: Interpretation and translation are the same (Gadamer, 1975, p. 365; Koller, 1987, p. 51; Bühler, 2002). This means that any act of interpretation, but especially those made of modalities and contexts foreign to the interpreter at a societal ‘level of culture’ (House, 2002) such as at a national level, are an act of translation. This means that researchers must bridge the chasm between linguistic systems and cultural differences that otherwise make their analysis null and void and the use of the artefact pointless, since it is these ‘differences’ that give cultures their ‘singularity’ (Deutsch, 1966, p. 75) and that should be the very focus of cultural translation (Bhabha, 1994). In short, Korean language and culture cannot be analysed through Western European scopes of reasoning (Hong, 2009; Kim and Kiaer, 2021; Kiaer, 2022; Kim, 2022, 2024). There is much documentation of this by researchers across the realms of translation (Bassnett and Lefevere, 1995), cultural studies (Bhabha, 1994), linguistics (Venuti, 2009; Kiaer, 2019), and film studies that deal with multimodality and discourses as we do here (Higson, 2000; Kim, 2006). Kaplan (1993, p. 9), for instance, comments on the limitations of analysing Chinese films, stating ‘cross-cultural analysis is difficult: It is fraught with danger. We are either forced to read works produced by the Other through the constraints of our own frameworks/ theories/ ideologies; or to adopt what we believe to be the position of the Other – to submerge our position in that of the imagined Other.’ Similarly, Matron (2010, p. 36) in analysis of Korean film as a West German researcher, states

‘[...] it is still important to keep in mind the position that is taken by the author. In this article, I drew a line connecting two movies from two very different cultures while always maintaining my own West German point of view. It is obvious that within the limited context of this study it is not possible to undertake a deeper comparison of the movies regarding diverging filmic and narrative traditions and the applicability of symbols specific to the respective culture.’

Willemen (2006, p. 35) argues that this gap between researcher and film must be accounted for, or otherwise conform to the cultural practices of the researcher:

‘If we accept that national boundaries have a significant structuring impact on national socio-cultural formations [...], this has to be accounted for in the way we approach and deal with cultural practices from “elsewhere.” Otherwise, reading a Japanese film from within a British film studies framework may in fact be more like a cultural cross-border raid, or worse, an attempt to annex another culture in a subordinate position by requiring it to conform to the readers’ cultural practices.’

If foodscaping, then, is to be undertaken by researchers non-native to a particular environment, it is crucial that processes are put in place to ensure cultural differences are taken into account such that specificity and primacy of the source culture is maintained. As per Willemen, it is unethical, not to mention pointless, to approach other cultures from one’s own cultural perspective. Following Eurocentric traditions to address Asian artefacts is akin to analysing a Western adaptation rather than the original text itself. K-SFDRS offers researchers the opportunity to apply a logical framework, grounded in the socio-pragmatics of Korean language and culture, where the gap between Korean and Western scopes predominantly resides, to Korean film so as to facilitate an unbiased understanding of the conventions and meanings made therein. SFDRS is developed to draw discourse structures based on how multimodality ‘makes sense.’ Hong argues that ‘Confucianism is the “common sense” that permeates all kinds of Korean social interactions’ and in order to understand Korean communication, it is Confucian reasoning that has been argued needs developing into a ‘functional and comprehensive tool’ (Hong, 2009, p. 5–7). Kim (2022) builds upon this by developing K-SFDRS with Korean Confucian socio-pragmatics, which constitute a significant, although not sole, influence on Korean interpersonal relations. For a more in-depth discussion of K-SFDRS and its importance to understanding Korean film, we direct readers to Kim (2024).

### 1.3.5 Event segmentation

Another aspect of SFDRS that we believe to hold particular potential for foodscaping analysis lies in the fact that the pre-existing SDRS methodology (Asher and Lascarides, 2003) was developed using Event Segmentation Theory, enabling its application to multimodal narrative discourses in film. Event segmentation theory applies not only to film, where there have been considerable studies, Song et al. (2021) and the earlier Zacks (2010) to name a few, but also to real-life information processing. In his delineation of the architecture of Event Segmentation Theory, Zacks (2020, p. 42) explains that ‘people can segment ongoing activity reliably with virtually no training. This seems to be picking up on something that is just a natural part of the observing activity.’ Some researchers have described this interpretation of events as a constant construction of narratives. Song et al. (2021), for example, state:

‘We make sense of our memory and others’ behaviour by constantly constructing narratives from an information stream that unfolds over time. Comprehending a narrative is a process of accumulating ongoing information, storing it in memory as a situational model, and simultaneously integrating it to construct



a coherent representation. Forming a coherent representation of a narrative involves comprehending the causal structure of the events, including the causal flow that links consecutive events or even a long-range causal connection that exists between temporally discontinuous events.'

Thus, the event segmentation inherent to the (K-)SFDRS methodology gives it potential to accurately analyse the information processing by actors in real-life interactions over food. The significance of segmentation specifically to cross-cultural interpretation has also been demonstrated by Kim (2024 forthcoming; Kim, 2022) in the Korean context. Kim (2022) found that socio-pragmatic primitives are used to infer meaning, and that there is an effect on event segmentation as a result when drawing SFDRS. Event-related potential (ERP) studies on the pragmatic processing of Korean honorifics in the brain have shown that when honorifics are misaligned, the N400 effect occurs, signalling pragmatic mismatches. N400 forms part of the common electrical brain activity observed in response to a variety of meaningful and potentially meaningful stimuli (Kiaer et al., 2022). The same N400 effect has been found to occur in response to modulations in film editing (Sanz-Aznar et al., 2023). As Kim (2022) argues, given that stimuli evoke specific functional reactions in an organ or tissue, there is potential for Korean socio-pragmatics (i.e., honorifics) to be determinants of how Koreans segment 'meaningful events,' and thus how discourses in Korean contexts – such as the Korean food environment we examine in this chapter – are understood by Koreans. A related study of Japanese language processing (Cui et al., 2022) is, since among East Asian languages Japanese is the most similar socio-pragmatically to Korean (Kiaer, 2018), 'still highly relevant here' (Kim, 2022). The study by Cui et al., which looks into functional magnetic resonance imaging, examines the neural correlates of honorific agreement processing mediated by socio-pragmatic factors in the Japanese language. This study demonstrates that socio-pragmatic factors, such as social roles and language experience, could be key influences in language processing, and demonstrates that social cues, such as social status, 'trigger computation of honorific agreement' (Cui et al., 2022, p. 1). This is to say that honorifics register as feature changes as encountered by actors in cultures that employ them. Therefore, the consideration of Korean socio-pragmatics by K-SFDRS fully takes account of their importance in event segmentation and, in turn, the discourses present and how they are understood.

## 1.4 Conclusion to the introduction

In short, the ability of K-SFDRS to engage with event segmentation according to uniquely Korean socio-pragmatic factors, identifying those which are most salient to a given interaction, gives it potential as a useful tool in understanding Korean communication and meaning-making within foodscapes. It incorporates both 'obvious' and 'obscure' meanings and lays them bare to both the Korean and non-Korean researcher and reader. As detailed above, this event segmentation is as relevant to spontaneous communication as it is to edited filmic discourse, thus making the principles of K-SFDRS transferable to videographic evidence of live foodscapes. As Kim (2022) states of K-SFDRS:

'The approach suits the ambiguity encountered in Korean socio-pragmatic communication in filmic discourse, which is often subtle and requires a socio-pragmatic sensitivity that is particular to Korean interpersonal relations, through which clear dependencies

of socio-pragmatic relevance between inferences can be identified. This means that the relational structure that holds the socio-pragmatic expressions can be as important for making inferences of segments as the verbal and non-verbal referents that they contain.'

## 2 Step-by-step guide

For this pilot, video data was collected at a barbeque restaurant in South Korea. The restaurant has both inside and outside seating areas and serves a variety of sliced raw meats (mostly pork and beef, but also chicken) that patrons then cook themselves on coal grills built into the tables. Among the four tables included in our recording, we have chosen table two, situated in the outside eating area, as our main example; the patrons of table two are two Korean men. This section of the paper uses this example to guide how the video recording and informal on-site transcriptive notes of a foodscape are transferred into a formal description.

### 2.1 Data collection

In order to apply (K-)SFDRS to foodscaping, an audiovisual recording is taken, turning the foodscape experience into a piece of audiovisual material, akin to the film scenes to which (K-)SFDRS is designed to be applied, from which modalities may be transferred into logical forms. The format of the video is not dissimilar from food 'vlogging,' in which a patron records themselves, their fellow diners, surroundings, dishes, and more in the process of eating, discussing the food, and interacting with people and objects in the wider food environment. Some of the stills from this recording have been inserted into the box notation of logical forms to aid in our demonstration, as Wildfeuer (2014) and Kim (2022, 2024) also do. Please note that in these stills, members of the public have had their faces blurred to maintain anonymity.

Since foodscaping intends to capture and analyse an experience, not just videographic materials, informal fieldnotes on multimodal happenings should also be made by the researcher during and after filming. This intends to capture as much of the modal interplay as possible (for example, noting down smells, tastes, sounds) beyond just that which is apparent from the audiovisual recording alone. These notes should include the point in time at which each happening occurs to facilitate their alignment with the final recording.

### 2.2 Informal transcription

The video footage is then reviewed several times and a comprehensive informal transcription completed. This notes down all multimodal instances, such as dialogue, activities, and gestures, in the order in which they occurred. This informal transcription brings together chronological instances apparent in the video footage as well as those from the researcher's fieldnotes. This process involves several rewatches of the footage and constant review and updating of the transcription; through multiple viewings of the footage, the consequential relations between the multimodal inferences become increasingly apparent.

The nature of the informal transcription process, being an early stage of analysis, requires the researcher to begin by writing down as



much modality as they can recognise; this list will be narrowed down to the most salient elements in the latter stages. The following excerpt (Excerpt 1) is taken from the informal transcription of the events at and surrounding Mr. A and Mr. B at table two. Some of these modalities occurred simultaneously. Only dialogue and gesture are presented in linear order. Further, the fieldwork researcher's informal transcription notes have been merged with this in order to include smell and touch where applicable.

EXCERPT 1 Hot but not humid (touch)

Rain (touch, sound, visual)

Sundown (visual, smell)

Stainless steel cups and water bowls (visual)

Grilled meat (smell and visual and sound)

Mr. A turns grilled meat (visual)

Meat sizzles (sound)

Fish from nearby stall selling seafood (smell and visual)

Mopeds slowly driving through (sound and visual)

Staff on stalls working on the street (sound and visual)

People shuffling by (sound and visual)

People holding umbrellas (visual)

People in suits passing through (visual, sound)

Delivery men passing through (visual, sound)

Servers in uniforms (visual)

Mr. A: “*Yeogi doenjangjjigae hana*” (여기 된장찌개 하나...) ‘One *doenjang jjigae* [soybean soup] over here... [indecipherable words]

Server: “*Honja deusigeyo?*” (혼자 드시게요?) ‘Is the *doenjang jjigae* [soybean soup] for one person?’

Mr. A: [Indecipherable words]

Mr. B: Eats from one of the side dishes

Mr. A: “*Ani, ani, ani*” (아니, 아니, 아니) ‘No, no, no’

Mr. A turns meat on the grill

Woman working at stall in background (visual)

Constant hum of unknown voices in the background (sound)

People holding umbrellas (visual)

Mr. A adds more meat to grill and turns it over (visual)

Meat sizzles (sound)

Server returns and serves *doengjang jjigae* [soybean soup]

Another server walks by

Mr. A: “*Nae nae*” (네 네) ‘Yes yes’

Mr. A passes tray back to the server

Mr. A: “*Oh igeot, igeot, igeot*” (Oh 이것, 이것, 이것) ‘Oh this, this, this’

Server receives scissors, bows, and leaves

Mr. A: “*Joesonghae kimchi jom...*” (죄송해 김치 좀...) ‘Excuse me, a little *kimchi* [fermented cabbage side dish] perhaps...’

Server: “*Oh yea yea nae*” (Oh 예 예 네) ‘Oh yes, yes, yes’

## 2.3 Identify socio-pragmatic expressions

The informal transcription is then examined for socio-pragmatic expressions. This helps to find direction in the analysis, since it is these expressions which provide clarity on the meaning of Korean multimodality (Kim and Kiaer, 2021; Kim, 2022, 2024), which we hypothesise in this pilot can then be linked to relevant Korean-specific discourses. Continuing with the same excerpt, the following example (Excerpt 2) shows the identification of socio-pragmatic expressions in the verbal language, non-verbal gestures, and in the ‘food language’ (Kiaer et al., 2024).

### EXCERPT 2

Mr. A: “*Yeogi doenjangjjigae hana*” (여기 된장찌개 하나...) ‘One *doenjang jjigae* [soybean soup] over here... [indecipherable words] (informal speech style, speaking to much younger man)

Server: “*Honja deusigeyo?*” (혼자 드시게요?) ‘Is the *doenjang jjigae* [soybean soup] for one person?’ (formal speech style, speaking to much older man, and customer)

Mr. A: [Indecipherable words]

Mr. A: “*Ani, ani, ani*” (아니, 아니, 아니) ‘No, no, no’ (informal speech style, one soup to share suggests intimacy between the patrons)

Mr. A turns meat on the grill (would be submissive if the second man at the table wasn’t also doing so and if gestures from both didn’t suggest equality)

People with umbrellas walk by (no socio-pragmatic value – atmospheric)

Mr. A adds more meat to grill and turns it over

Meat sizzles

Server returns and serves *doengjang jjigae* [soybean soup]

Another server walks by

Mr. A: “*Nae nae*” (네 네) ‘Yes yes’ (formal speech style)

Mr. A passes tray back to server. (man passes with one hand because he’s so much older and a customer, and server receives with two because he is much younger and a server – age is main factor)

Mr. A: “*Oh igeot, igeot, igeot*” (Oh 이것, 이것, 이것) ‘Oh this, this, this’ (informal speech style)

Server takes scissors, bows, and leaves. (men don’t bow back because they are customers and much older) [paying further attention to the socio-pragmatics of the man’s gestures, now his avoidance of eye contact can be contextualised and recognised as salient too, as can both the supporting of his right arm with his left hand when reaching across the men and his reluctance to do so often, be identified though partially concealed from view]

Mr. A: “*Joesonghae kimchi jom...*” (죄송해 김치 좀...) ‘Excuse me, a little *kimchi* [fermented cabbage side dish] perhaps...’ (informal speech style)

Server: “*Oh yea yea nae*” (Oh 예 예 네) ‘Oh yes, yes, yes’ (formal speech style)

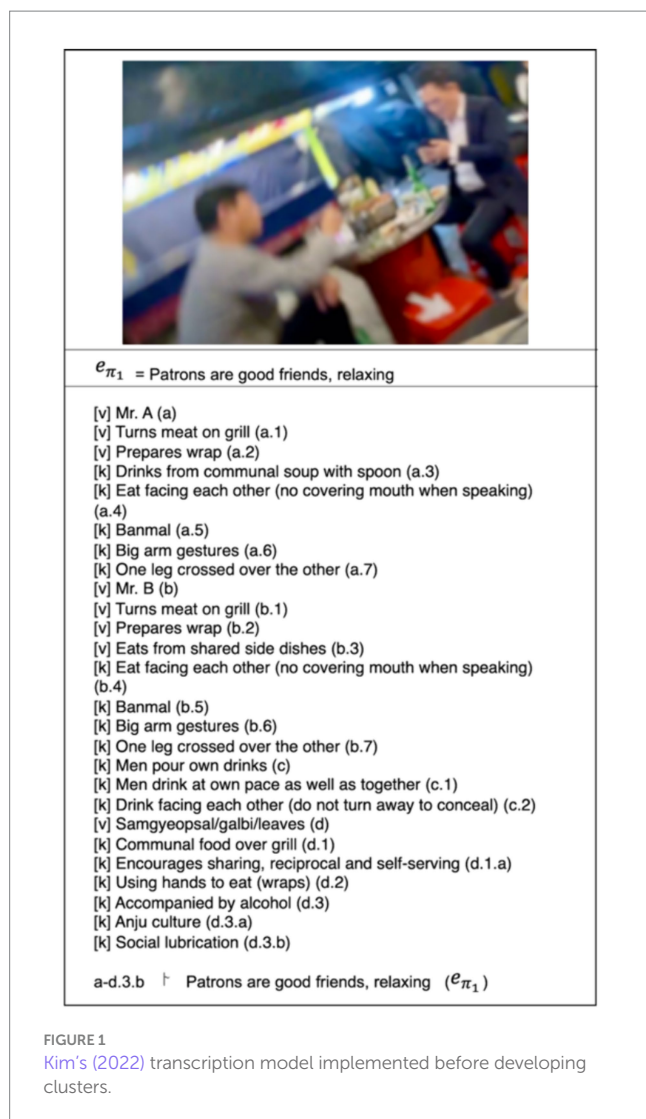
With the socio-pragmatic information taken from this informal transcription alone, we can ascertain the close intimate relationship of the two patrons dining at table two, which is important for discourses related to *anju* (안주, ‘food consumed customarily with alcohol’) and male bonding (Kiaer and Kim, 2021) as we shall go on to show, and the appropriate behaviour of the server (Figure 1).

## 2.4 Review and refine transcription

The referents seen in Figure 2 in logical forms  $e_{\pi_1}$ ,  $e_{\pi_1,a}$ ,  $e_{\pi_1,b}$ , and  $e_{\pi_1,c}$ , were inferred having collected the salient modalities from the informal transcription. Table 1 shows the modalities in their entirety assigned to Mr. A and Mr. B at table two. We have included referent and cluster form labels so that referents can easily be tracked to the graphical representations of logical forms that will follow.

## 2.5 Draw logical forms

Drawing the logical forms themselves, we follow Kim’s (2022, 2024) presentation, however, unlike Kim, Korean gestures and



speech styles have not been abbreviated in order to give as much clarity as possible in this early pilot. We do include the hierarchical relations between communicators, using < to indicate that the former person is junior to the latter (e.g., Person 1 < Person 2; person 1 is junior to person 2), and the reverse > (e.g., Person 1 > Person 2; person 1 is senior to person 2). This equation then specifies the type of seniority. (P) stands for 'position.' A senior in position can, and does in our analysis, refer to the senior position of a restaurant patron to service staff. (A) stands for 'age,' so if Person 1 were visibly senior in age, the hierarchical relation 'Person 1 > Person 2 (A)' would be attached to them within the logical form. In the logical form, salient modality is listed in the central area of the box and marked with [v] for 'visual,' [a] for 'audio,' or [k] for 'Korean' when Korean socio-pragmatic primitives are interpreted (e.g., speech styles, certain socio-pragmatic gestures, and the hierarchical relations previously mentioned) or other cultural communications. These 'referents' are also labelled, with the same labels that appear in the formula of the defeasible eventuality (the meaning interpreted from those multi-modal interactions), followed by the defeasible eventuality symbol, and finally the inferred meaning. Figure 3 shows a logical form in K-SFDRS format, created using the analysis presented in this chapter.

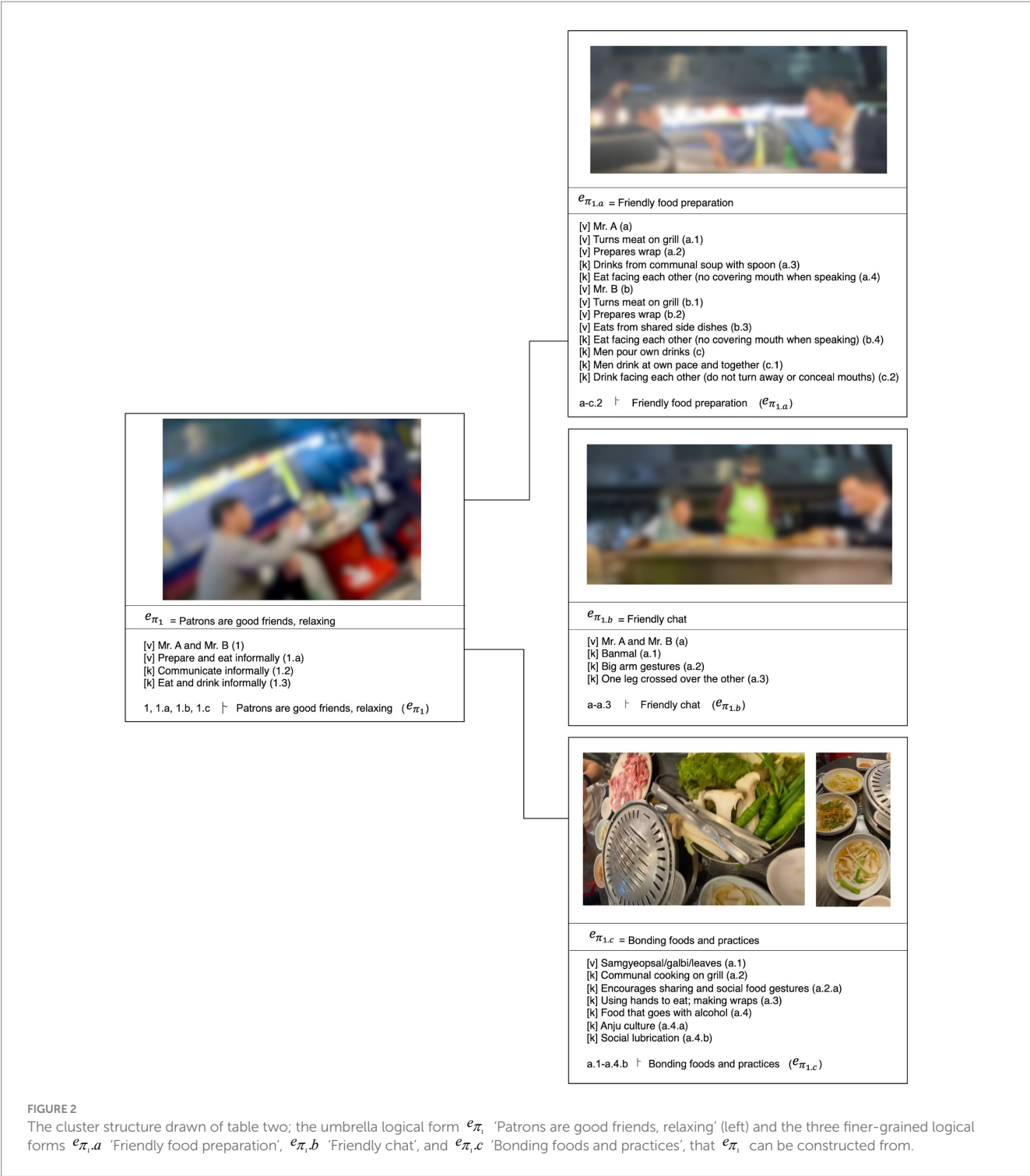
## 2.6 Organise logical forms and clusters

In the first instance, the above list of referents was considered within the scope of a single logical form, all of the components constructing a single meaning, but the referents are too numerous, and also offered separate, subtle contributions to the information inferable. To a degree, all referents share connections or commonalities, so it is possible to summarise (or rather *simplify*) what they amount to as a communication. This is because of the alignment of honorifics found in the verbal and non-verbal communication of Koreans (Brown, 2013; Kiaer and Kim, 2021; Kiaer, 2023), and the rigidity within which this is upheld, and the meaning potentials expressed if not (e.g., misalignment can be used to insult), that ultimately point to interpersonal relations and the nature of the interaction being observed.

It was decided early on that the pilot would need to be conducted without traditional discourse relation structures, and thus logical forms would be the focus. To understand how segments unfold in discourse, and apply discourse relations to the transcription, would require either a narrative or process to be observed or, where no narrative exists, a new conceptualisation of a narrative or process to replace it. It was thus determined that in the confines of this paper it was more prescient to focus on how the first granular level of K-SFDRS would work on independent instances of food consumption. However, this left another problem: several logical forms would be produced with an array of possibilities for interpretation, but they would be without the discourse relations which control said interpretation by highlighting whether they do or do not make sense. Fortunately, the process of creating logical forms itself granted the development of a method of organisation for the logical forms drawn from spontaneous, simultaneous multimodal discourse. The resulting structure is an umbrella logical form divided into various smaller logical forms, each of which represents a part of the main form. We term these 'cluster structures.' This approach does not require discourse relations, but rather transcribes how various meaning potentials are inferred by certain multimodal interactions and culminate ultimately in a particular aspect of the environment.

Figure 3 shows how  $e_{\pi_1}$  is divided, exposing a cluster of smaller logical forms, each of which possesses its own meaning potential and disciplinary routes through which it may be understood. See Figure 2 to view the complete logical forms of this cluster structure of  $e_{\pi_1}$ .

This development came about as it was observed that, when corroborative modalities could be identified working together, their unanimous meaning potential made them the salient modalities. Further, because of how socio-pragmatic rules limit the options for meaning potentials in Korean multimodal communication, as does socio-pragmatic alignment (Kiaer and Kim, 2021) and socio-pragmatic feature changes (Kim, 2022, 2024), meaning potentials were no longer elusive once these modalities could be identified. Logical forms could then be refined in order for them to 'make sense' both on Wildfeuer's (2014) and Kim's (2022, 2024) terms. Through this process of drawing logical forms, reviewing footage, and analysing alignment and consistency in expressions, not only is it possible to refine the referents listed in the box notations, but to refine the defeasible eventualities, and in some cases remove or merge logical forms altogether. Both Wildfeuer (2014) and Kim (2022, 2024) write on this process of revision and tweaking that occurs naturally when building (K-)SFDRS; the only difference here is simply that this occurred at the level of logical forms only. Each type of information



communicated by certain sets of modalities could then be separated into logical forms, and the logical forms clustered in the foodscape under analysis.

It is possible, with some consideration, to determine a set of categories that allow the divisions which facilitate clusters to take place, according to the corroboration of given modalities through which certain intentions or functions can be reasoned. For example, food gestures like 'drinking facing each other,' 'not concealing mouth when eating/drinking,' 'pouring own drink,' and 'drinking at own pace' are all expressions of informality and, consequentially, intimacy. The

same is true of similarly informal and intimate verbal language and non-verbal gestures such as speaking in 'banmal' (an informal speech style) and 'big arm gestures' that in formal or distant relations would be considered inappropriate. Eating meat with soju (a Korean spirit) is a social lubricant or setting for informal and intimate socialising – either to build intimacy in a relationship or for relationships in which this is already established. Both are linked by intimacy and informality, however, one a set of modalities (banmal, and pouring one's own drink, for example), while meat and soju are rather a well-suited setting for intimate and informal socialising and bonding. Please see

TABLE 1 Refined referents of informal transcription.

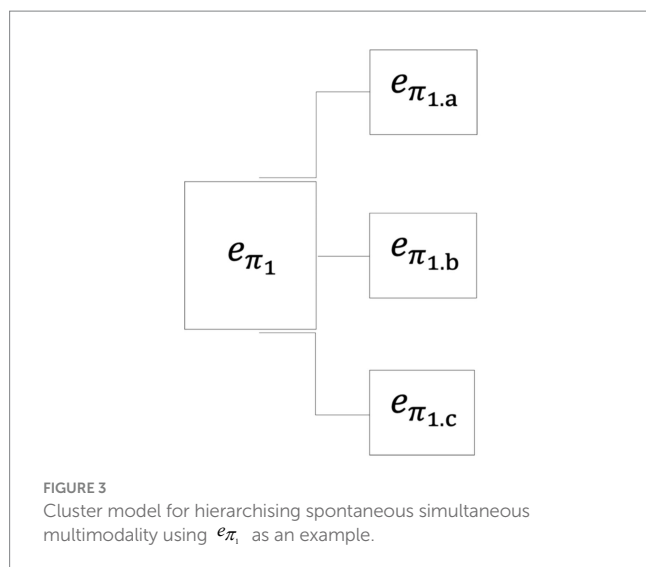
Patron	Referent	Referent label	Logical form cluster label
Mr. A	Turns meat on grill	(a.1)	$e_{\pi,a}$
	Prepares wrap	(a.2)	$e_{\pi,a}$
	Drinks from communal soup with spoon	(a.3)	$e_{\pi,a}$
	Eat facing each other (no covering mouth when speaking)	(a.4)	$e_{\pi,a}$
	<i>Banmal</i> (반말, 'half-talk')	(a.1)	$e_{\pi,b}$
	Big arm gestures	(a.2)	$e_{\pi,b}$
	One leg crossed over the other	(a.3)	$e_{\pi,b}$
Mr. B	Turns meat on grill	(b.1)	$e_{\pi,a}$
	Prepares wrap	(b.2)	$e_{\pi,a}$
	Eats from shared side dishes	(b.3)	$e_{\pi,a}$
	Eat facing each other (no covering mouth when speaking)	(b.4)	$e_{\pi,a}$
	Men pour own drinks	(c)	$e_{\pi,a}$
	Men drink at own pace together	(c.1)	$e_{\pi,a}$
	Drink facing each other (do not turn away or cover mouths)	(c.2)	$e_{\pi,a}$
	<i>Banmal</i>	(a.1)	$e_{\pi,b}$
	Big arm gestures	(a.2)	$e_{\pi,b}$
	One leg crossed over the other	(a.3)	$e_{\pi,b}$
Other	<i>Samgyeopsal</i> (삼겹살, sliced pork belly) / <i>galbi</i> (갈비, ribs) / leaves	(a.1)	$e_{\pi,c}$
	Communal cooking on grill	(a.2)	$e_{\pi,c}$
	Encourages sharing and social food gestures	(a.2.a)	$e_{\pi,c}$
	Using hands to eat, making wraps	(a.3)	$e_{\pi,c}$
	Food that goes with alcohol	(a.4)	$e_{\pi,c}$
	<i>Anju</i> culture	(a.4.a)	$e_{\pi,c}$
	Social lubrication	(a.4.b)	$e_{\pi,c}$

Kiaer and Kim (2021, 2024), Kiaer et al. (2024), and Kim (2022, 2024) where these expressions are covered extensively.

The umbrella logical form drawn up in this case was 'Patrons are good friends, relaxing.' This defeasible eventuality was inferred by both the socio-pragmatics of modalities employed by the two patrons, and their alignment and consistency. Both men used informal speech styles and gestures; even their postures, sitting with one leg crossed over the other, facing each other when downing shots of soju, neither

covering their mouth when eating and talking at the same time (Kiaer and Kim, 2021, 2024). Socio-pragmatic expressions out of alignment were purposeful and did not change this consistency. For instance, the men would take turns pouring drinks for each other on occasion, but this is a gesture that between friends is a way of making a fuss or showing care and not submissive (Kiaer and Kim, 2021), and combined with pouring drinks for oneself and one-handed, which would not be acceptable by a junior to a senior, does not nullify its





meaning. Another example is of how both men take turns tending to the meat on the grill; had one of the men been junior to the other then he would have solely or dominantly tended to the meat, and the elder tending to the meat would have been an exaggeration of care, which could be identified in its orchestration with other modalities such as back patting or saying ‘Eat meat!’ the Korean phrase *Geogi meogo!* (‘고기 먹어!’) that are also ways of showing care and therefore would support this inference.

This logical form can then be broken down into a ‘cluster’ in which finer grained meanings are defeasibly reasoned as segments constructing *Patrons are good friends, relaxing*; with modalities divided into groups according to their corroboration in meaning potential. As such, the notable long list of modalities found to be salient in the inference of *Patrons are good friends, relaxing* can be lessened in preparation for later discussion of discourses to be more specific.

Figure 2, the umbrella logical form  $e_{\pi_1}$ , and the cluster of logical forms  $e_{\pi_{1.a}}$ ,  $e_{\pi_{1.b}}$ , and  $e_{\pi_{1.c}}$  are the final logical forms in the cluster model developed during this pilot. They reflect the hierarchy of forms that combine simultaneously in spontaneous discourse. Eventuality labels (e.g.,  $e_{\pi_1}$ ) were applied with the umbrella logical form being a number (e.g., 1), and those that cluster to form said umbrella being labelled with a letter in addition to the respective number (e.g., 1.a, 1.b, 1.c, etc.). Further, the referents in umbrella logical forms are generalised (e.g., ‘informal communication’), while in the appropriate logical form from the cluster specified examples will be listed (e.g., ‘*banmal* speech style,’ ‘one-handed giving/receiving’). Each logical form within a given cluster was dedicated to one of the generalised modalities in the umbrella logical form which would be elaborated on within the logical form, and as such the logical form was labelled with the same referent label as the generalised modality, making the structure explicit.

### 3 Analysing the notation

Within the ‘friends eating’ cluster (the full version of the logical forms produced in the preceding step-by-step section), several themes

emerge that are ripe for interdisciplinary analysis. For the most part, these centre around the nature of Korean barbecue as a tool for socialisation and social lubrication, largely through alcohol drunk alongside the meal, as well as the reduction of politeness forms in interaction that fosters intimate friendship between social equals. Here we will give examples of avenues down which interdisciplinary researchers may go when expanding the SFDRS notation into a fully-fledged foodscaping analysis. We have labelled these avenues by the key theme that becomes apparent through the K-SFDRS analysis and suggest potential further avenues of investigation or elucidation scholars belonging to various disciplines may take.

#### 3.1 Friendship relations

Spoken Korean is highly mediated in terms of formality, putting a strong focus on and reinforcing a complex structure of social hierarchies (Kiaer and Kim, 2021). This permeates Korean speech in every situation, and determines the way in which Koreans speak to one another every minute of every day; for example, one must be aware of which register to use when talking to one’s boss, mother, older sibling, younger sibling, a friend one has just met, or a friend one meets every day and is close to, and how each one may differ from another in any given situation (Lee and Robert Ramsey, 2000, p. 267–272; Brown and Winter, 2019). Despite frequently entrenching hierarchies in a manner that keeps individuals emotionally distant, this feature of the Korean language is also an important tool in bond-building between individuals; deeming one’s relationship sufficiently intimate to transfer from a formal register to a more casual one can assert intimacy, friendship, and trust between individuals, provided it is instigated by the more senior of the pair or group and in an appropriate situation (Choo, 1999; Lee and Robert Ramsey, 2000).

A foodscape such as the one at the heart of our analysis, a barbecue restaurant of the most casual kind, by its nature encourages this kind of linguistic mediation from its patrons. The majority of the people eating in the restaurant on the evening in question were groups of close colleagues or young people in large friendship groups. In this example cluster, much of the salient multimodal referents are predicated on the manner in which they speak with one another, which can be categorised as ‘informal’ according to Korean speech registers. Through actions such as talking in ‘*banmal*’, gesturing in an animated way with one’s hands, and using typically ‘male’ registers of speech, the two friends are entrenching their close relationship with one another further throughout their meal (Choo, 1999; Lee and Robert Ramsey, 2000; Kiaer and Kim, 2021, 2024).

A researcher interested in the linguistic aspect of friends eating a meal may point towards this particular piece of notation in  $e_{\pi_{1.b}}$  as data-driven proof of the salience of said linguistic elements in the foodscape in question. Use of *banmal* and its consistency are shown in the notation to build significantly towards the resulting characterisation of the exchange as ‘Patrons are good friends, relaxing.’

Segment  $e_{\pi_{1.a}}$  carries the same themes of informality, but rather than just speech it demonstrates how the various referents contribute to the defeasible eventuality ‘Friendly food preparation.’ This demonstrates the gestural and food-specific dimensions of informal Korean linguistic analysis, thereby opening up the hypothesis that the communal cooking and eating processes in which patrons engage in this foodscape specifically encourages social bond-building. For

example, Mr. A takes food from a communal stew with his spoon; it is customary for a Korean meal to involve an individual bowl of soup or stew per diner, so in this instance the sharing of one bowl indicates the interlocutors' intimacy (Kiaer et al., 2024). Additionally, it is considered polite to cover one's mouth when speaking, an act of modesty and moderation when animated, in particular when laughing or smiling; an action in which neither of the men in this segment engage (Kiaer and Kim, 2021). Far from indicating impoliteness or rudeness to one another, however, the defeasible eventuality of this segment is 'Friendly food preparation.' Thus, we see how, with the correct relationship and level of intimacy, 'impolite' gestures such as those shown in  $e_{\pi,a}$  and  $e_{\pi,b}$  foster and reassert the close level of friendship between interlocutors; something which is facilitated by *anju* (Brown and Winter, 2019). This, then, indicates the foodscape in question, and perhaps, as an extension, barbecue restaurants in general, as a space which establishes an atmosphere in which informal behaviours are acceptable, even reaffirming the nature of the group as intimate friends.

### 3.2 Alcohol as a social lubricant

In c, c.1, and c.2 of  $e_{\pi,c}$  we can observe how soju is combined with the meal of Korean barbecue. A distilled alcoholic beverage which has been produced on the Korean peninsula from as early as the thirteenth century, soju is a very common beverage enjoyed by Koreans particularly as an accompaniment to communal meals such as Korean barbecue (Park, 2021). Although soju can range anywhere between about 10 and 50 percent ABV, lower alcohol soju ranging between 12 and 16 percent have become more frequent in the past few decades, making them akin to a wine in alcoholic terms (Park, 2014). Served to the table in cold bottles to be distributed into small glasses, frequently 'shot' sized with a 50 mL capacity, soju retains its cool temperature when being drunk in the hot atmosphere of a busy barbecue restaurant. The cool temperature and refreshing, slightly sweet, taste of soju makes it the perfect pairing to salty, fatty, caramelised barbecue meats (Yoon, 2022).

The importance of soju as a social lubricant, and the established norm in Korean culture of eating whilst drinking alcohol, further cements the success of soju and barbecued meat as a successful pairing (Ko and Sohn, 2018). In South Korea, the drinking of alcohol is most often combined with the eating of food, which may range anywhere from small snacks taken from packets, through a selection of cooked dishes, to full-blown meals featuring meat, stews, and rice; it is rare to drink at an establishment without some kind of food on the table to accompany it (Lee, 2011). The culture of *anju* has a long heritage in Korea, stemming centuries back into the dynastic periods of the peninsula, and is a tradition which endures amongst Koreans today, both young and old (Pettid, 2008).

In  $e_{\pi,a}$ , we particularly see how cultural roles associated with sharing alcohol, much like those of formal speech patterns and gestural behaviours, can be broken down and made more casual in order to assert close, typically male, friendships (Ko and Sohn, 2018). In  $e_{\pi,a}$ , we see both Mr. A and Mr. B pour soju from the communal bottle into their glass and drink at leisure. Where this might seem a relatively standard practice from a non-Korean perspective, it is in fact

notable that each man serves themselves and drinks as and when they choose. When socialising with others, particularly those of different social standings (such as seniority of work role, age, family position, etc.), it is common practice for the younger of the pair or group to serve others (often following order of seniority) from the shared bottle before they themselves are served by one of the other members of the group, and for everyone to drink at the same time, with juniors being sure to match pace with the more senior members of the group before the process is repeated again (Hines, 2022). It is also standard polite practice to cover one's mouth or turn away from one's interlocutors (especially seniors), when taking a sip from one's drink (Kiaer and Kim, 2024). In contrast to this, the interlocutors in  $e_{\pi}$  pour their own drinks, take sips of their drinks at their own pace, and drink facing one another without covering their mouths or turning away. Again, we see from the notation in  $e_{\pi,a}$  and  $e_{\pi,b}$  that such actions, rather than suggesting rudeness, contribute to the 'friendly' nature of their food preparation and consumption together.

Researchers of Korean socio-pragmatics may use the above analysis as an opportunity to explain, firstly, the relationships of Korean diners to one another as exemplified in the pouring and taking of drinks. The logical form could serve as a very useful piece of evidence in a comparative exercise with another, perhaps more formal, foodscape featuring both alcohol and a range of ages of participants to better exemplify this element of Korean culture. Food historians may also use the analysis to speak to the pairing of soju with barbecued meat, outlining their history and using the multimodal inferences as evidence to their combination by today's eaters. Sociologists and gender studies experts may further wish to analyse the interactions in light of notions of 'masculinity' associated with both the meat and the alcohol being consumed.

### 3.3 Eating 'correctly'

Several elements of the 'friends eating' cluster also illuminate the 'grammar' of a Korean meal and how this is observed by everyday Korean eaters in the context of a barbecue meal. Firstly, in  $e_{\pi,a}$  we observe patron Mr. B prepare a lettuce leaf wrap for eating, in which he takes a piece of meat that has been cooked on the communal grill and, holding an open lettuce leaf in the palm of his hand, wraps it up alongside sauce (assumedly *ssamjang* (쌈장)), *kimchi*, and vegetarian *banchan* (반찬, small side dishes customarily served for free alongside a meal) of his choosing into a bite-sized piece (the main ingredients for this are shown in  $e_{\pi,c}$ ). It is a common practice when eating Korean-style barbecue to take green leaves, which is often lettuce but can be any leafy green or a combination thereof and use it to wrap up a small piece of meat in combination with condiments (often a spicy or plain fermented bean paste), *banchan*, *kimchi*, or rice before eating in a single bite. When asking a Korean why they prefer to eat their meat as *ssam* (쌈), literally meaning 'wrapped,' they will most likely simply respond 'because it tastes better' (Lee, 2016). But this 'simple' aspect of taste can be traced back once again to traditional Korean preferences for balance not just across a meal or within a recipe, but in each individual bite. Owing to traditions of Traditional Korean Medicine and Neo-Confucianism, achieving a balance not only of flavours but also of colours, temperatures, and textures, is very important in the construction of a 'proper' Korean meal (Pettid, 2008). Indeed, whilst the *banchan*, condiments, meats, and rice of a Korean

barbecue meal demonstrate this balance, they are served in distinct dishes and are cooked separately, making it difficult to enjoy them all together at once, thereby properly savouring their combination and, therefore, the balance of flavours they together achieve. This is where *ssam* come in – the leafy green is used as a utensil in which the balanced flavours of each element of the barbecue meal can be combined. Researchers with a host of disciplinary backgrounds, including History, Traditional Korean Medicine, and Philosophy, would be valuable in fully unravelling the cultural and historical contexts relevant to this line of enquiry.

During  $e_{\pi,a}$  and  $e_{\pi,b}$ , we also see the *banchan*, or ‘side dishes,’ served to every table of diners for free as part of the meal, regardless of their other orders. The *banchan* served to the tables of the present foodscape are several without being numerous; diners are offered a dish of spicy dressed bean sprouts, shredded cabbage with a sesame dressing, vinegar-dressed greens, and a small side of *kimchi*. These three dishes are very commonly served at barbecue restaurants such as this one. *Banchan*, an integral part of any Korean meal, regardless of the price point, mealtime, or situation, serves a multitude of purposes in the Korean meal, but most importantly they balance the meal’s flavour profile, inject crucial vitamins and minerals to the diet, and add colour to the table (Kim, 2020). *Banchan* is served in small dishes, and it is the norm that diners may ask for a top-up of any given side dish as they eat (Kiaer et al., 2024) – in the foodscape analysis, Mr. A can be seen asking for an additional portion of *kimchi*. In other foodscapes the authors have observed, however, *banchan* (for which there is a long history in Korea stemming from royal court cuisine), are offered instead in the form of a self-serve *banchan* bar (Yeong, 2021). This allows patrons to select their own *banchan* from an array of options (the authors have observed restaurants offering anywhere between three and nine different side dishes and varieties of *kimchi*) and bring it to their table, asking only that they leave none at the end of their meal beyond reasonable leftovers. This change in long-standing format of meal presumably stems from economic considerations on the part of the restaurant, particularly following the COVID-19 pandemic and the resulting loss in sales (Lee and Koo, 2023). From an environmental perspective also, although we suspect this to be secondary to economic concerns from the perspective of restaurant owners, offering a lower volume of *banchan* which patrons may ask to top up should they wish more cuts down on food waste, which is increasingly a hot topic in Korean and global society. Indeed, as with many other elements of this foodscape, we see the realities on a local scale of a changing and shifting global South Korea reckoning with new concerns that appear to clash with traditional practices and expectations. Scholars interested in economics, public health, policy, and sustainability might all fruitfully follow this line of enquiry to develop a foodscaping analysis.

## 4 Conclusion and further research

In this paper we have outlined the need for a formal logic that can be applied to the multimodal analysis of foodscaping, proposed the use of the logical forms of (K-)SFDRS as a potential tool for this, and demonstrated (K-)SFDRS’s box notation style on a recorded live foodscape, as well as offering analysis of the multimodal strands of analysis and associated discourses that are revealed and structured by the notation. This has been completed using the example of an evening

foodscape at a barbeque restaurant in South Korea, which we have anonymised here.

The box notations and their analysis presented here lends a formal method to the understanding of a foodscape, featuring both local and non-local actors, in terms of diverse modalities, enabling the researchers to attach different disciplinary strands and contexts to further explain how certain inferences are produced in the foodscape in question: (1) the foodscape was recorded through informal transcription and audiovisual recording; (2) the data, categorised into ‘clusters’ based on different actors in the foodscape (in this case, tables of diners), was transformed into logical forms using the (K-)SFDRS model of box notation, through which; (3) the clusters were systematically analysed to produce a commentary on the salient multimodal elements that contribute to the inferences of given actors in each cluster, drawing on interdisciplinary discussions to give important context and background to explain how said inferences are ascertained from the associated multimodal influences.

The process of drawing equations, in which various modalities in the food environment possess meaning-making potential and combine to create various inferences, was found beneficial for identifying how certain modes interact to produce inferences, and the many multimodal discourses present in a food environment at a given time; not to mention the multifarious interdisciplinary roots of meaning-making processes that perhaps are not clear on initial inspection to the casual observer. This is because the modalities revealed by the (K-)SFDRS notation to be salient in creating certain meanings can be traced culturally and historically, thereby revealing the value of interdisciplinary analysis in understanding meaning-making processes in food spaces. Under scrutiny, this information then reveals connotations that accompany the local experience, in this case of *samgyeopsal* eateries in metropolitan spaces in South Korea, and specifically of the case study featured in this paper.

The process detailed in this paper has revealed advantages (both evident and potential), as well as drawbacks (both inherent and open to adjustment) in the use of logical forms of (K-)SFDRS for foodscape analysis. Here we outline these in greater detail.

### 4.1 Strengths

The ‘cluster’ system developed for the collection of segments into useful groupings has aided in the structure and logical discussion of the foodscape. The clusters allow researchers to focus on certain actors at any one time, essentially using primarily a nonlinear approach so as to deal with associated narratives without being overly dependent on chronology. We propose that cluster structures are ultimately catering formally to the incredibly complex ‘multidimensional maps’ that Doxiadis (2010, p. 81) describes: ‘Narratives flow linearly in time, yet they mediate between worlds that are largely nonlinear: both the world of action, with its manifold possibilities, and our mental models of it are like complex, multidimensional maps, representing not just objects but also relations, in webs of immense connectivity. Narratives by contrast, are like specific paths taken through these worlds — partial, linear views of nonlinear environments.’

The multimodal notation of (K-)SFDRS, through both audiovisual recording and fieldnotes, takes into account all modalities and brings them together as equal participants in meaning-making. This ensures



that it is not only the most 'apparent' modalities that are taken into account; the researcher must consider every utterance, regardless of modality, as of equal potential importance until the informal transcription is narrowed down to salient modalities.

Using the (K-)SFDRS notation as a basis for discussion and analysis, drawing discussions based on the relevant discipline as they appear, has proved useful in establishing a neutral foundation from which to explore the multifaceted avenues of the foodscape. Further, by identifying pertinent actions (i.e., those which contribute the most to a given inference), the notation ensures that conversation springs only from the multimodal occurrences which result in genuine meaning-making. Thus, the method acts as a great base upon which to build layers of interdisciplinary analysis only when pertinent to the inference at hand, and without relying on a given discipline as 'prime' or as the foundation for analysis.

The formal logic of (K-)SFDRS notation contributes to the understanding of how meaning is made through certain multimodal interactions in a way that can be explained to locals and non-locals alike. Our analysis, in particular the table demonstrating knowledge sources, illustrates how using (K-)SFDRS notation enables researchers to clearly align a stimulus with the mode in which it occurs, the inference in which it results, and the knowledge required for actors to reach said inference. This explicitly joins the dots of the meaning-making processes within any given foodscape and for any given actor, clearly exposing the roots of inferences, sparking discussions on the similarities and differences therein, and providing opportunities in which to explain the interdisciplinary background where pertinent.

The production of logical forms facilitates the in-depth review of an experienced foodscape (through both audiovisual recording and defeasible notes made at the time) and their transferal into, firstly, a list of multimodal occurrences and, secondly, logical forms based on their importance and interaction to create meaning. During this paper, we have found the very action of producing logical forms on the part of the researcher to be helpful in attributing inferences (which otherwise rely solely on the researcher's own intuition) to data-based evidence. As such, the use of (K-)SFDRS in the process of analysing a foodscape helps the researcher to double check one's own assumptions and intuitive thoughts, cross-referencing them with the salient moments evidenced through the process of producing (K-)SFDRS logical forms.

The appearance of final logical forms in box notation (as well as intermediary stages such as lists of multimodal occurrences if relevant) and umbrella clusters helps the reader of foodscaping outputs to better understand and contextualise the analysis in several ways. Firstly, the description of a foodscape, considering its rich multimodal nature, is a difficult task in the context of an academic paper; word limits are often strict, and the literary language required for description is often undesirable. (K-)SFDRS notation has the potential to overcome this by laying out key multimodal instances (either/both chronological and thematic) for the reader in a concise, objective manner. Secondly, in the process of reading foodscaping analysis, readers are directed repeatedly to different instances of meaning-making as demands the flow of the paper's argument. As such, logical forms and their labelling are useful as a reference point to which readers may return in order to fully contextualise the referents discussed at points in the argumentation. Lastly, much like the researcher themselves, logical forms enable readers to cross-check author assertions with the multimodal evidence to which they pertain, again ensuring that conclusions do not rely exclusively on author intuition.

## 4.2 Limitations

(K-)SFDRS, owing to its main purpose as film notation, relied most heavily upon audiovisual data as the main source of analysis. Although our implementation of an informal transcription, which foregrounds sensory data, as well as ensuring the researchers personally experience the foodscape whilst recording, has helped to incorporate further non-audiovisual, multimodal aspects into the analysis, further experimentation may be needed to develop a more robust method of incorporating multimodal evidence into the notation.

Following on from Strength 1, we have found clustering a useful tool. Though, as we previously stated and expected given that discourse relations were not able to be applied to logical forms simply, the non-linear aspect has taken precedence. Whilst chronology is not essential in each inference, we would like to develop clusters to engage more actively with linear processes in food consumption, preparation, and etiquette. We believe this would better demonstrate the foodscape as it happens and allow readers to contextualise multimodal inferences better in the process of following the final analysis.

## 4.3 Further research

The present study only makes the very first inroads into the potential use of (K-)SFDRS methodology in the pursuit of foodscaping projects. Whilst we believe that this methodological paper has proved the potential of the methodology and its strengths and weaknesses, there are several key avenues down which we believe future research could profitably travel to further explore and verify its usefulness.

This paper, due to constraints on length and its methodological nature, does not utilise (K-)SFDRS in the context of a full and proper foodscaping analysis. As such, it cannot definitively evidence the usefulness of the methodology. We propose further studies entirely within the foodscaping methodology which make use of (K-)SFDRS in the initial analysis so as to further investigate its applicability.

We additionally propose that future research employ (K-)SFDRS in foodscaping projects which incorporate authors from various disciplinary backgrounds so as to make the most of the multimodal nature of the logical forms and their ability to serve as a 'jumping off point' for multiple disciplinary avenues. Based on disciplinary background, our expertise as authors lends itself to linguistics, multimodality, cross-cultural perspectives, Korean culture, history, and food; as such, these are the topics with which our example analyses most heavily engage. Should researchers who are specialists in Korean agriculture, politics, or economy, for example, employ this methodology, their results might accordingly identify contributions of other modalities to other logical forms. Rather than a hindrance, we see this as mirroring the interdisciplinary nature of foodscaping as an endeavour; interdisciplinary teams of researchers are a benefit, if not an essential, to the aims of foodscaping, and thus the simultaneous use of the (K-)SFDRS method on the same multimodal inferences by researchers with different disciplinary backgrounds is a necessity for the full and proper analysis of a foodscape.

We have discussed how (K-)SFDRS, being developed as a method for film analysis (and therefore relying on strong cinematic narratives), must be altered to fit spontaneous discourse in which purposeful narrative does not exist. As such, discourse relations were not



forthcoming in the foodscape analysis presented here, resulting in the original innovation of the cluster notation. However, as outlined in limitation 2, we believe that the logical form clusters would be best complimented with a chronological or relation-based method specifically designed to suit spontaneous discourse. This would enable the true-to-life chronology of the foodscape to be better conveyed to the reader, and, in turn, allow relations to be drawn between occurrences in a manner more akin to the original SFDRS methodology. For this, a further novel notation style must be developed; this is an important task that we must leave to future studies.

As scholars with backgrounds in Korean studies, and using the K-SFDRS system developed by Kim (2022, 2024), this paper necessarily focuses on the Korean context for its example analyses. Part of the benefits of K-SFDRS over SFDRS evidenced in this paper is its specificity to the Korean language and socio-pragmatic nuances, both in reference to native actors and to the interactions of non-native actors within the context of a South Korean foodscape. As such, we believe that the development of specific local SFDRS systems is crucial to this method's success in further international endeavours. We encourage our colleagues in Japanese studies to consider J-SFDRS, Chinese studies to consider C-SFDRS, *etcetera*, until a wealth of international logical forms can be drawn from to facilitate nuanced and accurate cross-cultural foodscaping analyses.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## References

- Asher, N., and Lascarides, A. (2003). *Logics of conversation*. Cambridge: Cambridge University Press.
- Bassnett, S., and Lefevere, A. (1995). "General editors' preface" in *The Translator's invisibility: a history of translation*. ed. L. Venuti (London: Routledge).
- Bateman, J. A., and Wildfeuer, J. (2014). A multimodal discourse theory of visual narrative. *J. Pragmat.* 74, 180–208. doi: 10.1016/j.pragma.2014.10.001
- Bhabha, H. (1994). *The location of culture*. London: Routledge.
- Bohnemeyer, J., Enfield, N. J., Essegbey, J., Ibarretxe-Antunano, I., Kita, S., Lüpke, F., et al. (2007). Principles of event segmentation in language: the case of motion events. *Language* 83, 495–532. doi: 10.1353/lan.2007.0116
- Brown, L. (2013). 'Mind your own esteemed business': sarcastic honorifics use and impoliteness in Korean TV dramas. *J. Politeness Res.* 9, 159–186. doi: 10.1515/pr-2013-0008
- Brown, L., and Winter, B. (2019). Multimodal Indexicality in Korean: 'doing deference' and 'performing intimacy' through nonverbal behavior. *J. Politeness Res.* 15, 25–54. doi: 10.1515/pr-2016-0042
- Bühler, A. (2002). "Translation and interpretation" in *Translation studies: perspectives on an emerging discipline*. ed. A. Riccardi (Cambridge: Cambridge University Press).
- Calway, N., Loh, K., and Robert, W.-C. (2025). *Foodscaping the Seoul metropolis: foodscaping Korea*. London: Bloomsbury.
- Choo, M. (1999). Teaching language styles of Korean. *Korean Lang. Am.* 3, 77–95.
- Cui, H., Jeong, H., Okamoto, K., Takahashi, D., Kawashima, R., and Sugiura, M. (2022). Neural correlates of Japanese honorific agreement processing mediated by socio-pragmatic factors: an fMRI study. *J. Neurolinguistics* 62:101041. doi: 10.1016/j.neuroling.2021.101041
- Deutsch, K. (1966). *Nationalism and social communication: an inquiry into the foundations of nationality*. Massachusetts: Massachusetts Institute of Technology Press.
- Doxiadis, A. (2010). Narrative, rhetoric, and the origins of logic. *Storyworlds J. Narrat. Stud.* 2, 79–99. doi: 10.5250/storyworlds.2.1.79
- Gadamer, H.-G. (1975). *Truth and method*. New York: Seabury Press.
- Gerhardt, C., Frobenius, M., and Ley, S. (2013). "Culinary linguistics: the Chef's special" in eds. C. Gerhardt, M. Frobenius and S. Ley (Amsterdam/Philadelphia: John Benjamins Publishing Company).
- Higson, A. (2000). "The limiting imagination of National Cinema" in *Cinema and nation*. eds. M. Hjort and S. MacKenzie (London: Routledge), 63–87.
- Hines, Nick. (2022). What is soju? Everything you need to know about Korea's national drink. Available at: <https://vinepair.com/articles/soju-koreas-national-drink/#soju-how>
- Hong, J. O. (2009). *A discourse approach to Korean politeness: towards a culture-specific confucian framework*. Nottingham: Nottingham Trent University.
- House, J. (2002). "Universality versus culture specificity in translation" in *Translation studies: perspectives on an emerging discipline*. ed. A. Riccardi (Cambridge: Cambridge University Press), 92–110.
- Kaplan, E. A. (1993). "Melodrama / subjectivity / ideology: western melodrama theories and their relevance to recent Chinese cinema" in *Melodrama and Asian Cinema* (Cambridge: Cambridge University Press), 9–28.
- Kiaer, J. (2018). *The Routledge course in Korean translation*. London: Routledge.
- Kiaer, J. (2019). "Translating invisibility: the case of Korean-English literary translation" in *Translation and literature in East Asia*. 1st ed. Eds. K. Jieun, G. Jennifer and L. Xiaofan Amy (London: Routledge).
- Kiaer, J. (2022). *Pragmatic particles: findings from Asian languages*. London: Bloomsbury Academic.
- Kiaer, J. (2023). *The language of Hallyu: more than polite*. London: Routledge.
- Kiaer, J., and Kim, L. (2021). *Understanding Korean film: a cross-cultural perspective*. 1st Edn. London: Routledge.
- Kiaer, J., and Kim, L. (2024). *Embodied words: a guide to Asian non-verbal gestures through the lens of film*. Oxon: Routledge.
- Kiaer, J., Kim, L., and Calway, N. (2024). *The language and food through the lens of East Asian film and dramas*. London: Routledge.
- Kiaer, Jieun, Lee, I., and Brown, L. (2022). *An ERP study on the pragmatic processing of Korean honorifics and politeness*. (Forthcoming).

## Author contributions

LK: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. NC: Data curation, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Kim, S. K. (2006). Renaissance of Korean National Cinema as a terrain of negotiation and contention between the global and the local: analysing two Korean blockbusters, Shiri (1999) and JSA (2000). Available at: <https://www.semanticscholar.org/paper/-Renaissance-of-Korean-National-Cinema--as-a-of-%3A-Kim/e03e2065951c51e9ed89ba78bf6c30fa48c6e87f# citing-papers>
- Kim, Eric. (2020). A spread worthy of loyalty. The New York Times. Available at: <https://www.nytimes.com/2020/09/28/dining/banchan-recipes.html>
- Kim, L. (2022). *A theory of multimodal translation for cross-cultural viewers of south Korean film*. University of Oxford.
- Kim, L. (2024). *Interpreting Korean film discourse*. London: Routledge.
- Kim, L., and Kiaer, J. (2021). "Conventions in how Korean films mean" in *Empirical multimodality research*. eds. J. Wildfeuer, J. Pflaeging and J. Bateman (Berlin: De Gruyter), 237–258.
- Ko, S., and Sohn, A. (2018). Behaviours and culture of drinking among Korean people. *Iran J. Public Health* 47, 47–56.
- Koller, W. (1987). *Einführung in Die Ub" Ersetzungswissenschaft*. Heidelberg: Quelle und Meyer.
- Lee, Cecelia Hae-Jin. (2011). Food and drinks the Korean way. Available at: <https://www.latimes.com/food/la-xpm-2011-may-26-la-fo-anju-20110526-story.html>
- Lee, Jinjoo. (2016). How to enjoy Ssam (Korean lettuce wraps) with different greens. Available at: <https://kimchimari.com/leaves-used-for-korean-lettuce-wraps-ssam-ssambap/>
- Lee, C., and Koo, Y. (2023). Analyzing sales of the Korean restaurant franchise during the COVID-19 pandemic with the mixed-effects model approach. *PLoS One* 18:e0293147. doi: 10.1371/journal.pone.0293147
- Lee, I., and Robert Ramsey, S. (2000). *The Korean language*. Albany: State University of New York Press.
- MacKendrick, N. (2014). Foodscape. *Contexts* 13, 16–18. doi: 10.1177/1536504214545754
- Matron, A. (2010). Transferability of cultural meanings: a case study on contemporary German and South Korean cinema. *Literature & Aesthetics*, 20, 26–37.
- Park, Eun-Jee. (2014). Koreans Looking for Weaker Soju. Available at: <https://koreajoongangdaily.joins.com/news/article/article.aspx?aid=2997481>
- Park, H. (2021). *Soju: A global history*. Cambridge: Cambridge University Press.
- Pettid, M. J. (2008). *Korean cuisine: an illustrated history*. London: Reaktion Books.
- Sanz-Aznar, J., Bruni, L. E., and Soto-Faraco, S. (2023). Cinematographic continuity edits across shot scales and camera angles: an ERP analysis. *Front. Neurosci.* 17, 1489–1517. doi: 10.3389/fnins.2023.1173704
- Song, H., Park, B.-y., Park, H., and Shim, W. M. (2021). Cognitive and neural state dynamics of narrative comprehension. *J. Neurosci.* 41, 8972–8990. doi: 10.1523/JNEUROSCI.0037-21.2021
- Venuti, L. (2009). Translation, intertextuality, interpretation. *Roman. Stud.* 27, 157–173. doi: 10.1179/174581509X455169
- Wildfeuer, J. (2012). *Coherence in film and the construction of logical forms of discourse: a formal-functional perspective*. der Universität Bremen, Bremen: Universität Bremen.
- Wildfeuer, J. (2014). *Film discourse interpretation. Towards a new paradigm for multimodal film analysis*. New York: Routledge.
- Willemen, P. (2006). "The nation revisited" in *Theorising national cinema*. eds. V. Vitali and P. Willemen (London: BFI and Palgrave Macmillan), 29–43.
- Yeong, . (2021). Korean side dishes | the story of Korea's many banchan. Available at: <https://creatrip.com/en/blog/8577>
- Yoon, Sojung. (2022). Paring Korean food with booze (2): soju and beer. Available at: <https://www.korea.net/NewsFocus/FoodTravel/view?articleId=222951#:~:text=%22Be cause%20soju%20has%20a%20high,alcoholic%20beverages%2C%22%20she20added>
- Yu, Janna. (2017). Korean barbecue: the essence of social dining. Available at: <https://cornellsun.com/2017/03/25/korean-barbeque-the-essence-of-social-dining/>
- Zacks, J. M. (2010). The brain's cutting-room floor: segmentation of narrative cinema. *Front. Hum. Neurosci.* 4:168. doi: 10.3389/fnhum.2010.00168
- Zacks, J. M. (2020). "Event segmentation theory and the segmentation of visual events" in *Ten lectures on the representation of events in language, perception, memory, and action control* (Leiden, Boston: BRILL), 38–54.

# Frontiers in Communication

Investigates the power of communication across  
culture and society

A cross-disciplinary journal that advances our  
understanding of the global communication  
revolution and its relevance across social,  
economic and cultural spheres.

## Discover the latest Research Topics

See more →

### Frontiers

Avenue du Tribunal-Fédéral 34  
1005 Lausanne, Switzerland  
[frontiersin.org](https://frontiersin.org)

### Contact us

+41 (0)21 510 17 00  
[frontiersin.org/about/contact](https://frontiersin.org/about/contact)



### Frontiers in Communication

