

Cognitive infocommunications

Edited by

Anna Esposito, Gennaro Cordasco, Carl Vogel and Péter Baranyi

Published in

Frontiers in Computer Science

Frontiers in Psychology



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 978-2-8325-1903-5
DOI 10.3389/978-2-8325-1903-5

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

Cognitive infocommunications

Topic editors

Anna Esposito — University of Campania 'Luigi Vanvitelli, Italy
Gennaro Cordasco — University of Campania Luigi Vanvitelli, Italy
Carl Vogel — Trinity College Dublin, Ireland
Péter Baranyi — Széchenyi István University, Hungary

Citation

Esposito, A., Cordasco, G., Vogel, C., Baranyi, P., eds. (2023). *Cognitive infocommunications*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-1903-5

Table of contents

04	Editorial: Cognitive infocommunications Anna Esposito, Gennaro Cordasco, Carl Vogel and Péter Baranyi
11	Automatic Estimation of Multidimensional Personality From a Single Sound-Symbolic Word Maki Sakamoto, Junji Watanabe and Koichi Yamagata
27	Developing an Error Map for Cognitive Navigation System Atsushi Ito, Yosuke Nakamura, Yuko Hiramatsu, Tomoya Kitani and Hiroyuki Hatano
41	Speech Pauses and Pronominal Anaphors Costanza Navarretta
54	Experiment in a Box (XB): An Interactive Technology Framework for Sustainable Health Practices m. c. schraefel, George Catalin Muresan and Eric Hekler
77	What Does the Body Communicate With Postural Oscillations? A Clinical Investigation Hypothesis Andrea Buscemi, Santi Scirè Campisi, Giulia Frazzetto, Jessica Petrilligieri, Simona Martino, Pierluca Ambramo, Alessandro Rapisarda, Nelson Mauro Maldonato, Donatella Di Corrado and Marinella Coco
84	A Canonical Set of Operations for Editing Dashboard Layouts in Virtual Reality Tarek Setti and Adam B. Csapo
93	Robot Concept Acquisition Based on Interaction Between Probabilistic and Deep Generative Models Ryo Kuniyasu, Tomoaki Nakamura, Tadahiro Taniguchi and Takayuki Nagai
107	An Analysis of Personalized Learning Opportunities in 3D VR Ildikó Horváth
119	Unifying Physical Interaction, Linguistic Communication, and Language Acquisition of Cognitive Agents by Minimalist Grammars Ronald Römer, Peter beim Graben, Markus Huber-Liebl and Matthias Wolff



OPEN ACCESS

EDITED AND REVIEWED BY

Kostas Karpouzis,
Panteion University, Greece

*CORRESPONDENCE

Anna Esposito

✉ Anna.Esposito@unicampania.it

Carl Vogel

✉ vogel@cs.tcd.ie

SPECIALTY SECTION

This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

RECEIVED 22 December 2022

ACCEPTED 31 January 2023

PUBLISHED 27 February 2023

CITATION

Esposito A, Cordasco G, Vogel C and Baranyi P
(2023) Editorial: Cognitive infocommunications.
Front. Comput. Sci. 5:1129898.
doi: 10.3389/fcomp.2023.1129898

COPYRIGHT

© 2023 Esposito, Cordasco, Vogel and Baranyi.
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: Cognitive infocommunications

Anna Esposito^{1,2*}, Gennaro Cordasco^{1,2}, Carl Vogel^{3*} and Péter Baranyi^{4,5}

¹Department of Psychology, University of Campania "Luigi Vanvitelli", Naples, Italy, ²International Institute for Advanced Scientific Studies (IIASS), Vietri sul Mare, Italy, ³Trinity Centre for Computing and Language Studies, School of Computer Science and Statistics, Trinity College Dublin, The University of Dublin, Dublin, Ireland, ⁴Széchenyi István University, Győr, Hungary, ⁵Eötvös Loránd Research Network, Győr, Hungary

KEYWORDS

CogInfoCom, cognitive infocommunications, learning, knowledge representation, robotics, speech, interaction, navigation

Editorial on the Research Topic Cognitive infocommunications

1. Introduction

Cognitive Infocommunications is a nascent focal point of study at the intersection of multiple disciplines, ergo—a multi-discipline. The research community includes anthropologists, designers, economists, educationalists, engineers, gerontologists, linguists, philosophers, psychologists, psychotherapists, mathematicians, statisticians, roboticists, and more. Researchers who are embedded in these many disciplines, whether well-established or at an early stage, are united by interest in the questions that arise from wonder about enhancements to human cognitive capabilities. As childhood learning (and learning more generally) is an instance of human cognitive enhancement, it is no wonder that educationalists, who study methods by which human learning may be enhanced, contribute equally to cognitive infocommunications with technologists who seek to develop new technologies that may extend human cognitive capabilities and cognitive scientists who wish to understand how humans think and interact in whatever setting humans find themselves.

This volume presents papers that develop further works that were presented at the 2020 instance of an annual conference on cognitive infocommunications.¹ As such, this volume presents a coherent snapshot of current activity and recent advances in this area of focus. These works do not exhaust the scope of the area, but are representative of some key themes; matters of: human embodiment, knowledge representation, learning, speech interaction, navigation, and frameworks for virtual interaction. This list reiterates the fact that in studying methods, including the provision of new technologies, for enhancing human cognitive capabilities, it is necessary to continue studying the scope of extant human cognitive behaviors and how humans respond to new technologies.

¹ Apart from the annual conference in the field some journals regularly feature work from cognitive infocommunications researchers (see [Sallai, 2012](#)), including with special issues of scholarly journals (see [Baranyi et al., 2012](#); [Baranyi and Csapo, 2014](#); [Fujita and Baranyi, 2014](#); [Baranyi, 2020](#); [Katona, 2022](#)) and books (see [Baranyi et al., 2015](#); [Klempous et al., 2019, 2022](#)).

We write to describe contributions from the following (listed in the order in which the finale versions were published): [Sakamoto et al.](#) (22 April 2021), [Ito et al.](#) (4 May 2021), [Navarretta](#) (13 May 2021), [schraefel et al.](#) (26 May 2021), [Buscemi et al.](#) (16 June 2021), [Setti and Csapo](#) (12 July 2021), [Kuniyasu et al.](#) (3 September 2021), [Horváth](#) (20 September 2021), and [Römer et al.](#) (27 January 2022). It is noteworthy that some of the papers are provided via *Frontiers in Computer Science*, others via *Frontiers in Psychology*, and all through the section attending to human-media interaction. We situate these contributions by first providing more details of the nature of CogInfoCom, via its history (§2). We then provide a synthetic review of the works published here (§3). We give our current view on the priorities for next steps and longer term challenges in this multi-discipline (§4). Finally, we conclude (§5).

2. An overview of CogInfoCom

At the outset, we described cognitive infocommunications as “nascent,” and as this is fitting in the area of study that is roughly 15 years old, but however it which draws on fields that have persisted for millennia. Initial conceptualizations of the field are provided by [Baranyi and Csapo](#) (2010, 2012). The essence is as described at the outset—attention to technologies that enhance human cognitive capabilities. An example is in electronic calculators: humans do not need them in order to perform arithmetic computations; however, they enable a greater volume and accuracy of computation to an extent that humans have come to depend upon them. One may think of calculators as an example of successful cognitive infocommunications technology. [Baranyi et al.](#) (2014) reviewed the state of progress in the field at an earlier stage in its development, and it can be seen that the essential focus has remained constant, although the technologies and problems addressed vary. At the time of writing this introduction (December 2022), using “CogInfoCom” as a search term within IEEEExplore returns 1,057 results and, about 5,030 results within GoogleScholar, about 5,030 results. We mention these facts in order to give an indication of the scale of this scientific community.

[Vogel and Esposito](#) (2019, 2020) have analyzed the nature of successful cognitive infocommunication technologies, arguing that any such technology is one that solves problems faced by humans in private spheres (such as in thought) and public spheres (such as in interaction). With this schema, they present “technologies,” such as human language and clothing, as well-entrenched advances that have become assimilated into the concept of humanity, such assimilation constituting evidence of their success. Language is perhaps even more useful for thinking than for communication, but it is certainly adapted to problems in both spheres; clothing, for example, provides individual protection against varying climate conditions (in the private sphere), but also provides a means to signal membership in groups (in the public sphere). This philosophical analysis helps to understand that new technologies, seemingly developed for their own sake, are frequently explored by researchers in this field. Part of the exploration is about the technology in itself, but this inherently addresses the problems that such technologies solve. Part of the exploration is about the manner in which people interact with and potentially embrace the technology, and thus user evaluations are sometimes invoked.

As an example additional to language as an abstract technology that has been assimilated into the collective concept of humanity, “mathability” may also be noted: consider, for example, the conceptual advance created by adding the notion of “zero” to integer arithmetic. Accordingly, mathematics as a technology and technologies for supporting mathematical reasoning provide a robust thread of research within cognitive infocommunications ([Baranyi and Gilányi, 2013](#); [Török et al., 2013](#); [Szi and Csapo, 2014](#); [Chmielewska et al., 2016](#)). Money is a comparable technology, and the advent of cryptocurrencies reveals that it is a category of technology still undergoing innovation; it is therefore no surprise that economics is a contributing discipline in cognitive infocommunications ([Matarazzo et al., 2016](#); [Esposito et al., 2017](#); [Vogel et al., 2018](#)). The works in this volume are, without exception, attuned to the question of enhancing human cognitive capabilities. They vary in their approach.

3. Synthetic view of the works in this volume

As indicated, some of the research themes in cognitive infocommunications include human embodiment, knowledge representation, learning, speech interaction, navigation, and frameworks for virtual interaction. We describe the papers presented in this special issue in relation to our view of those papers and themes.

3.1. Human embodiment

[Buscemi et al.](#) study human posture, and the role of fascia, the tissue that surrounds muscles, bones, and organs in the human body. A test involving the application of mechanical pressure to a standing human accompanied by a change in sensory input, in particular, closure of eyes, is revealing of specific tissue tensions—stiffness or loss of elasticity in fascia tissue. This, in turn, may provide insight into therapies. A protocol for further study is offered. We see as the main idea here of relevance is a conceptually simple technology that may provide information to an individual or medical professionals about physical imbalances experienced by an individual, thereby enabling remedy. Others who have contributed to cognitive infocommunications have also been investigating correlating features sensed from human body postures into diagnostics of motion ([Granata et al., 2013](#)), health and wellbeing, ([Steiner and Kertesz, 2012](#); [Macik, 2018](#); [Mykhailova et al., 2018](#)) and ability ([Toma et al., 2012](#)).

3.2. Knowledge representation

[Sakamoto et al.](#) address the manner in which words involve sound-symbolism cross-linguistically, but with different categories in each language. They focus on Japanese words and their encoding of personality traits using sound symbols, and through sourcing multiple independent judgements of personality traits associated with symbols, offer a means of concise personality classification

labeling, reliable in the sense of that associations with personality traits are shared. This work supports the explicit representation of knowledge implicit in the underlying sound symbolism about of human personality traits.

Kuniyasu et al. examine the formation of concepts through robot analysis of multi-modal stimuli—images and words. They provide a new latent Dirichlet allocation model that supports unsupervised learning of relevant feature individuation (as opposed to prior specification of or pre-training feature extraction) and cross-modal inference from words to images. The paper provides the background to and details of the model, highlighting its successes and areas where it may be improved. Additional stimulus modalities are identified as a priority in future work. We see this research as contributing a perspective on the co-temporality of object perception and labeling of those objects by cognitive agents.

Römer et al. provide a theory of environment-situated interaction among agents expressed using formal language theory and lambda calculus, adopting minimalist grammars and utterance-meaning transducers. Agent lexicons are learned through interaction, and a reinforcement learning regime supports this. A closed-world simulation is provided to illustrate the details of the approach. We view this work as an analysis of an extant cognitive infocommunications technology (natural language, in this case) and the dynamics of its use.

Knowledge representation is central to cognition, and it follows that that representation systems, knowledge acquisition, and reasoning processes are studied by many within this multi-discipline. To mention a few examples, we draw attention to the needs of representing spatial information (Karimipour and Niroo, 2013), computer system design (Shakhnov et al., 2013), business processes (Szmodics, 2015), medical decision making (Minutolo et al., 2014), knowledge acquisition (Cao et al., 2013), knowledge provenance tracking (Mittal et al., 2018), and, of course, cognitively informed representation systems (Dhuieb et al., 2013; Savić et al., 2017).

3.3. Learning

The paper by schraefel et al. addresses user evaluation of heuristics for maintaining health and wellbeing through a paradigm of systematic self-experimentation. They seek to establish a technology for interaction that facilitates users in developing an active base of knowledge, skills, and practice in health and wellbeing maintenance that ultimately makes the technology inessential. In this context, three studies are presented. The first examined within a small group ($n = 12$) whether the self-experimentation paradigm could be well-received and lead to behavioral change that persists beyond the experiment (it did). The second study involved an application in support of the self-experimentation paradigm, and achieved the desired effects. The third study presented an experiment that compared standard care and the experimental paradigm, finding positive impacts from its support for engagement and emphasis on self-awareness and reflection.

Horváth studied human learning in a 3D virtual reality environment, attending to systematic differences in learner styles. Investigations revealed the importance of tuning task specifications

to learner styles. An affordance of the virtual reality environment in supporting non-intrusive direction was identified as particularly effective.

We have characterized the form of learning instantiated by acquisition of knowledge within computer systems as a problem of knowledge representation. However, human learning is also relevant. Wilcock and Yamamoto (2015) address the use of robots in human language learning, and this is representative of work that assesses the integration of technology in human learning processes (Horváth, 2016; Biró et al., 2017; Sik-Lanyi et al., 2017). Researchers in the area also study learning processes in the abstract (Gogh and Kovari, 2018). More detailed positioning of human learning supported by technology within cognitive infocommunications is reviewed by Kovari (2018); Kovari et al. (2020).

3.4. Speech interaction

Navarretta analyzes qualities of speech (pauses and stress) in relation to the antecedents of third person genderless singular pronouns (e.g. “it” in English or “det” in Danish), as recorded in a Danish language map-task corpus and Danish corpus of first encounters between people. Contrasts between abstract and individual referents as candidate antecedents are noted. Stress on the pronouns favors abstract antecedents; unstressed pronouns favor individual antecedents. Silent and filled pauses prior to the pronoun are more common with abstract referents than individual referents; this is interpreted as a signal of the relative difficulty of the concept denoted.

As noted above, it has been argued that language is an early cognitive infocommunications technology. Speech is a modality for expression of language, and CogInfoCom researchers dwell on very many aspects of speech, such as—speech recognition, speaker recognition, speech synthesis, speech as an information source for clinical diagnosis, speech in relation to other co-linguistic signals, such as gesture, and so on (Esposito and Esposito, 2011; Meena et al., 2012; Szaszák and Beke, 2012; Vicsi et al., 2012; Kiss et al., 2013; Marinozzi and Zanuy, 2014; Siegert et al., 2015; Hunyadi et al., 2016; Navarretta, 2016, 2017; Kovács et al., 2017; Beke, 2018; de Velasco Vázquez et al., 2019; Esposito et al., 2020).

3.5. Navigation

Ito et al. provide and validate a method for constructing maps of error in positioning, such as is relevant needed to guided navigation. They analyze error sources (e.g., features of global positioning satellites or obstructions by buildings or natural landscape features) and provide a means of estimating the extent and location of error. The methods are validated. The notion of an error map developed is a powerful quantification of uncertainty. It is easy to imagine its adoption within autonomous navigation (e.g., yielding control back to a human driver within probable high-error regions), as well as in the guided navigation currently used by drivers and pedestrians: it would be helpful to a pedestrian following a dot on a map in an unknown city to know that the pedestrian is on the cusp of a high-error region.

Representations of spatial knowledge afforded by maps and supports for navigation in real and virtual environments are a concern of many within CogInfoCom (Furuyama and Niitsuma, 2013; Fixova et al., 2014; Jämsä et al., 2014; Karimipour et al., 2015; Török et al., 2017; Török and Török, 2018, 2019; Berki, 2019, 2020b; Sudár and Csapó, 2019).

3.6. Frameworks for interaction

Setti and Csapo reason that 3D virtual reality environments deter user embrace acceptance through lacking because of the lack of 2D controls that are familiar from other technology settings. They describe their work in enhancing a popular 3D virtual reality environment (MaxWhere) with 2D user controls that may be available in the 3D space for users to manipulate that space. This entails conceptualizing an inventory of 2D operations small enough to avert excess choice, but large enough to be effective, and providing instantiations of the operations. User testing reveals the approach to require of users an acceptable amount of time and control over the environment. It can be foreseen that this approach will increase uptake of this and other 3D virtual reality environments.

Horváth, whose work was outlined above, also addressed advantageous features of 3D virtual reality environments. Many other researchers also study virtual reality from a cognitive infocommunications perspective — more than 150 such works have been published in the annual conference series. Approaches typically include developing, evaluating, and improving VR experiences (Lóska, 2012; Köles et al., 2014; Lógó et al., 2014; Persa et al., 2014; Lanyi et al., 2015; Berki, 2018a,b, 2020a; Tolgay et al., 2019; Ishii et al., 2020; Esposito et al., 2021) and scrutiny of the value of VR in a range of applications (Qvist et al., 2016; Nascivera et al., 2018; Vér, 2018; Budai and Kuczmán, 2019; Markopoulos et al., 2019; Berki, 2020b; Csapó et al., 2020).

3.7. Synthesis

These works touch on areas of focus that are current within cognitive infocommunications: human embodiment, knowledge representation, learning, speech interaction, autonomous navigation, and frameworks for interaction. As noted in the last section, the papers included here sometimes address matters at the intersection of a number of these areas. For example, Römer et al. present work that explores knowledge representation, learning, and interaction. The work of Horváth is at the intersection of learning and virtual reality environments that support interaction. The investigation by Navarretta of the information value of pauses and stress on certain pronouns is relevant to both speech interaction and human knowledge representation of the contrast between reference to abstract concepts and individuals. The contribution of Setti and Csapo has a focus on virtual reality applications, but also optimization of representation of knowledge in providing a means of specifying interactions within such an application. While Ito et al. are focused on the problem of

navigation, the representation of loci of uncertainty transcends the navigation application domain. The approach to emergent knowledge representation used by Kuniyasu et al. impinges on learning, and with has wide applicability. The work of schraefel et al. proposes the Experiment in a Box protocol of short self-experiments to help participants carry out and reflect upon guided experiences with fundamental health and wellbeing practices—the approach is designed to help participants connect, or “insource,” these knowledge, skills, and practices with how they feel; similarly, Buscemi et al. focus on embodiment, but also toward on identifying paths to improved health and wellbeing. The results reported by Sakamoto et al. are in the context of representing personality traits, but thereby draw attention to a dimension of lexical choice in human linguistic interaction that merits deeper scrutiny.

4. Projections for the future

The works included in this volume grew out of presentations given at the 2020 conference on cognitive infocommunications—given the speed of some technology, this is already some time ago. Nonetheless, the themes that we have highlighted are relatively timeless. While these works present advances, none presents offer the final word.

Given a focus on technologies that may enhance human cognitive capabilities, we think that an open problem in the field is in establishing limits. Time and again in disciplines that are supported by mathematical reasoning, profoundly important developments have taken the form of establishing limits, such as the limits to learn ability of context-free languages on the basis of positive examples alone (Gold, 1967) (explicitly important to Römer et al.), or the limits of computability. It remains an open question whether human cognitive capabilities are infinitely amenable to enhancement, or whether there are limits overall or in the domain of individual categories of cognition: it will advance the field to establish formal proof, either way.

We anticipate ongoing analysis of how humans respond, individually and in interaction, to the technologies they face and potentially adopt, just as we anticipate continuing exploration of new technologies that address problems that humans face individually and collectively.

We think that among these new technologies will be a focus on privacy preservation while (perhaps seemingly paradoxically) enabling greater free constructive expression.

5. Conclusion

As indicated above, we do not see efforts in advancing cognitive infocommunications to have concluded. We hope this paper to adds to the field by establishing how recent works have advanced the field and by articulating what we deem to be the most important open questions.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Acknowledgments

We express gratitude to Anna Sudár for her continuing support of events within cognitive infocommunications as well as research contributions in the area; we also thank the many researchers, including the researchers whose papers are included in this volume, who have provided individual and collective advances to this field of study. Further, we thank the reviewers whose efforts have sharpened and enhanced the papers presented in this volume (in alphabetical order): Tiziano A. Agostini, Borbála Berki, Dubravko Culibrk, Mike Fraser, Tsutomu Fujinami, Jozsef Katona, Javad Khodadoust, Attila Kovari, Maria Koutsombogera, Farhan Mohamed, György Molnár, Mihoko Niitsuma, Juan Sebastian Olier, Rubén San-Segundo, Miriam Sturdee, Erzebet Toth, and Matej M. Tusak.

References

- Baranyi, P. (2020). Special issue on cognitive infocommunications. Theory and applications - guest editorial. *Infocommun. J.* 12:1. Available online at: https://www.infocommunications.hu/documents/169298/4652492/InfocomJ_2020_1_0_GuestEditorial_Baranyi.pdf
- Baranyi, P., and Csapo, A. (2010). "Cognitive infocommunications: CogInfoCom," in *2010 11th International Symposium on Computational Intelligence and Informatics (CINTI)* (Budapest), 141–146.
- Baranyi, P., and Csapo, A. (2012). Definition and synergies of cognitive infocommunications. *Acta Polytech. Hung.* 9, 67–83. Available online at: https://uni-obuda.hu/journal/Baranyi_Csapo_33.pdf
- Baranyi, P., and Csapo, A. (2014). Special issue on multimodal interfaces in cognitive infocommunication systems. *J. Multimodal User Interfaces* 8, 119–120. doi: 10.1007/s12193-014-0155-2
- Baranyi, P., Csapo, A., and Sallai, G. (2015). *Cognitive Infocommunications (CogInfoCom)*. Cham: Springer International Publishing.
- Baranyi, P., Csapo, A., and Varlaki, P. (2014). "An overview of research trends in CogInfoCom," in *IEEE 18th International Conference on Intelligent Engineering Systems (Tihany: INES)*, 181–186.
- Baranyi, P., and Gilányi, A. (2013). "Mathability: emulating and enhancing human mathematical capabilities," in *IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 555–558.
- Baranyi, P., Hashimoto, H., and Sallai, G. (2012). Cognitive infocommunications. *J. Adv. Comput. Intellig. Intellig. Inform.* 16, 283–367. doi: 10.20965/jaciii.2012.p0283
- Beke, A. (2018). "Forensic speaker profiling in a Hungarian speech corpus," in *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 000379–000384.
- Berki, B. (2018a). "Better memory performance for images in MaxWhere 3D VR space than in website," in *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 000281–000284.
- Berki, B. (2018b). "Desktop vr and the use of supplementary visual information," in *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 000333–000336.
- Berki, B. (2019). "Sense of presence in MaxWhere virtual reality," in *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Naples), 91–94.
- Berki, B. (2020a). "Level of presence in max where virtual reality," in *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Mariehamn), 000485–000490.
- Berki, B. (2020b). "Navigation power of MaxWhere: a unique solution," in *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Mariehamn), 000509–000514.
- Biró, K., Molnár, G., Pap, D., and Szuts, Z. (2017). "The effects of virtual and augmented learning environments on the learning process in secondary school," in *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Debrecen), 000371–000376.
- Budai, T., and Kuczmanski, M. (2019). "A multi-purpose virtual laboratory with interactive knowledge integration," in *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Naples), 529–532.
- Cao, M., Li, A., Fang, Q., and Kröger, B. J. (2013). "Growing self-organizing map approach for semantic acquisition modeling," in *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 33–38.
- Chmielewska, K., Gilányi, A., and Łukasiewicz, A. (2016). "Mathability and mathematical cognition," in *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Wrocław), 245–250.
- Csapó, Á. B., Sudár, A., Berki, B., Gergely, B., Berényi, B., and Czabán, C. (2020). "Measuring virtual rotation skills in MaxWhere," in *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Mariehamn), 000587–000590.
- de Velasco Vázquez, M., Justo, R., Zorrilla, A. L., and Torres, M. I. (2019). "Can spontaneous emotions be detected from speech on TV political debates?," in *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Naples), 289–294.
- Dhuiab, M. A., Laroche, F., and Bernard, A. (2013). "Toward a cognitive based approach for knowledge structuring," in *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 407–412.
- Esposito, A., Amorese, T., Cuciniello, M., Riviello, M. T., Esposito, A. M., Troncone, A., et al. (2021). Elder user's attitude toward assistive virtual agents: the role of voice and gender. *J. Ambient Intellig. Human. Comput.* 12, 4429–4436. doi: 10.1007/s12652-019-01423-x
- Esposito, A., and Esposito, A. M. (2011). "On speech and gestures synchrony," in *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*, eds A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, and A. Nijhol (Berlin; Heidelberg), 252–272.
- Esposito, A., Esposito, A. M., Esposito, M., Scibelli, F., Cordasco, G., Vogel, C., et al. (2017). "How traders' appearances and moral descriptions influence receivers' choices in the ultimatum game," in *2017 International Conference on Tools with Artificial Intelligence* (Boston, MA), 409–416.

Conflict of interest

PB is co-founder of the MaxWhere virtual reality (VR) platform, which is analyzed by two of the papers in this Research Topic; however, those two papers were independently edited and reviewed. Further, no commercial gains followed, and results reported are transferable to other 3D VR platforms.

The authors declare that the research was conducted in the absence of any other 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.

- Esposito, A., Raimo, G., Maldonato, M., Vogel, C., Conson, M., and Cordasco, G. (2020). "Behavioral sentiment analysis of depressive states," in *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Mariehamn), 000209–000214.
- Fixova, K., Macik, M., and Mikovec, Z. (2014). "In-hospital navigation system for people with limited orientation," in *2014 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom)* (Vietri sul Mare), 125–130.
- Fujita, H., and Baranyi, P. (2014). Special issue on: knowledge-bases for cognitive infocommunications systems (kbics). *Knowl. Based Syst.* 71, 1–2. doi: 10.1016/j.knsys.2014.08.017
- Furuyama, T., and Niitsuma, M. (2013). "Building a multi-resolution map for autonomous mobile robot navigation in living environments," in *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 261–266.
- Gogh, E., and Kovari, A. (2018). "Metacognition and lifelong learning : A survey of secondary school students," in *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 000271–000276.
- Gold, E. M. (1967). Language identification in the limit. *Inform. Control* 16, 447–474.
- Granata, C., Salini, J., Ady, R., and Bidaud, P. (2013). "Human whole body motion characterization from embedded Kinect," in *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 133–138.
- Horváth, I. (2016). "Innovative engineering education in the cooperative VR environment," in *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Wroclaw), 000359–000364.
- Hunyadi, L., Szekrényes, I., and Kiss, H. (2016). "Prosody enhances cognitive infocommunication: Materials from the HuComTech corpus," in *Toward Robotic Socially Believable Behaving Systems - Volume 1 : Modeling Emotions*, eds A. Esposito and L. C. Jain (Cham: Springer International Publishing), 183–204.
- Ishii, S., Luimula, M., Yokokubo, A., and Lopez, G. (2020). "VR dodge-ball: application of real-time gesture detection from wearables to exergaming," in *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Mariehamn), 000081–000082.
- Jämsä, J., Luimula, M., Heinonen, T., and Leskinen, A. (2014). "Utilizing point cloud data in the cognitive car navigation," in *2014 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom)* (Vietri sul Mare), 575–576.
- Karimipour, F., and Niroo, A. (2013). "Agent-based modelling of cognitive navigation: How spatial knowledge is constructed and used," in *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 17–22.
- Karimipour, F., Niroo, A., and Alinaghi, N. (2015). "Generalizing route descriptions based on the user's spatial knowledge: considering the user's spatial cognitive information in communicating route instructions," in *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 89–94.
- Katona, J. (2022). Research directions of applications of cognitive infocommunications (CogInfoCom). *Appl. Sci.* 12, 8589. doi: 10.3390/app12178589
- Kiss, G., Sztahó, D., and Vicsi, K. (2013). "Language independent automatic speech segmentation into phoneme-like units on the base of acoustic distinctive features," in *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 579–582.
- Klempous, R., Nikodem, J., and Baranyi, P. Z., editors (2019). *Cognitive Infocommunications, Theory and Applications*. Cham: Springer.
- Klempous, R., Nikodem, J., and Baranyi, P. Z., editors (2022). *Accentuated Innovations in Cognitive Info-Communication*. Cham: Springer.
- Köles, M., Hámornik, B. P., Lógó, E., Hercegfí, K., and Tóvölgyi, S. (2014). "Experiences of a combined psychophysiology and eye-tracking study in VR," in *2014 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom)* (Vietri sul Mare), 391–395.
- Kovács, A., Winkler, I., and Vicsi, K. (2017). "EEG correlates of speech: Examination of event related potentials elicited by phoneme classes," in *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Debrecen), 000115–000120.
- Kovari, A. (2018). "Coginfocom supported education : a review of CogInfoCom based conference papers," in *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 000233–000236.
- Kovari, A., Katona, J., Heldal, I., Demeter, R., Costescu, C., Rosan, A., et al. (2020). "Relationship between education 4.0 and cognitive infocommunications," in *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Mariehamn), 000431–000436.
- Lanyi, C. S., Tolgyesi, S. M., Szucs, V., and Toth, Z. (2015). "Wheelchair driving simulator: Computer aided training for persons with special need," in *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Gyor), 381–384.
- Lógó, E., Hámornik, B. P., Köles, M., Hercegfí, K., Tóvölgyi, S., and Komlódi, A. (2014). "Usability related human errors in a collaborative immersive VR environment," in *2014 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom)* (Vietri sul Mare), 243–246.
- Lóska, Á. (2012). "BCI and virtual collaboration," in *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)* (Kosice), 89–94.
- Macik, M. (2018). Cognitive aspects of spatial orientation. *Acta Polytech. Hung.* 15, 149–167. doi: 10.12700/APH.15.5.2018.5.9
- Marinozzi, S., and Zanuy, M. F. (2014). "Digital speech algorithms for speaker de-identification," in *2014 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom)* (Vietri sul Mare), 317–320.
- Markopoulos, E., Lauronen, J., Luimula, M., Lehto, P., and Laukkanen, S. (2019). "Maritime safety education with VR technology (MarSEVR)," in *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Naples), 283–288.
- Matarazzo, O., Pizzini, B., Greco, C., and Carpentieri, M. (2016). "Effects of a chance task outcome on the offers in the ultimatum game: the mediation role of emotions," in *7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Wroclaw), eds Baranyi, P., Klempous, Z., Mikovec, Z., and Pieska, S., 295–300.
- Meena, R., Jokinen, K., and Wilcock, G. (2012). "Integration of gestures and speech in human-robot interaction," in *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)* (Kosice), 673–678.
- Minutolo, A., Esposito, M., and De Pietro, G. (2014). "An ontology-based fuzzy approach for encoding cognitive processes in medical decision making," in *2014 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom)* (Vietri sul Mare), 397–402.
- Mittal, U., Gedeon, T., and Caldwell, S. (2018). "Managing knowledge provenance using blockchain," in *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 000241–000246.
- Mykhailova, P., Zubkov, S., Antonova-Rafi, J., and Khudetskyy, I. (2018). "Application of the photoplethysmography technique to complex wireless diagnostic the functional state of the human body," in *2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T)* (Kharkiv), 308–312.
- Nascivera, N., Alfano, Y. M., Annunziato, T., Messina, M., Iorio, V. S., Cioffi, V., et al. (2018). "Virtual empathy : The added value of virtual reality in psychotherapy," in *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 000321–000326.
- Navarretta, C. (2016). "Predicting an individual's gestures from the interlocutor's co-occurring gestures and related speech," in *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Wroclaw), 000233–000238.
- Navarretta, C. (2017). "Prediction of audience response from spoken sequences, speech pauses and co-speech gestures in humorous discourse by Barack Obama," in *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Debrecen), 000327–000332.
- Persa, G., Török, Á., Galambos, P., Sulykos, I., Kecskés-Kovács, K., Czizler, I., et al. (2014). "Experimental framework for spatial cognition research in immersive virtual space," in *2014 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom)* (Vietri sul Mare), 587–593.
- Qvist, P., Trygg, N. B., Luimula, M., Peltola, A., Suominen, T., Heikkinen, V., et al. (2016). "Demo: Medieval gastro box – utilizing VR technologies in immersive tourism experiences," in *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Wroclaw), 000077–000078.
- Sallai, G. (2012). The cradle of the cognitive infocommunications. *Acta Polytech. Hung.* 9, 171–181. Available online at: https://uni-obuda.hu/journal/Sallai_33.pdf
- Savić, S., Gnjatović, M., Mišković, D., Tasevski, J., and Maček, N. (2017). "Cognitively-inspired symbolic framework for knowledge representation," in *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Debrecen), 000315–000320.
- Shakhnov, V., Makarchuk, V., Zinchenko, L., and Verstov, V. (2013). "Heterogeneous knowledge representation for VLSI systems and MEMS design," in *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), pages 189–194.
- Siebert, I., Bock, R., Wendemuth, A., Vlasenko, B., and Ohnemus, K. (2015). "Overlapping speech, utterance duration and affective content in HHI and HCI – an comparison," in *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Gyor), 83–88.
- Sik-Lanyi, C., Shirmohammadi, S., Guzsvinecz, T., Abersek, B., Szucs, V., Van Isacker, K., et al. (2017). "How to develop serious games for social and cognitive competence of children with learning difficulties," in *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Debrecen), 000321–000326.
- Steiner, H., and Kertesz, Z. (2012). "Effect of therapeutic riding on gait cycle parameters and behavioural skills of autistic children," in *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)* (Kosice), 109–113.

- Sudár, A., and Csapó, Á. B. (2019). "Interaction patterns of spatial navigation in VR workspaces," in *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Naples), 615–618.
- Szaszák, G., and Beke, A. (2012). "Automatic prosodic and syntactic analysis from speech in cognitive infocommunication," in *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)* (Kosice), 377–382.
- Szi, B., and Csapo, A. (2014). "An outline of some human factors contributing to mathability research," in *2014 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom)* (Vietri sul Mare), 583–586.
- Szmodics, P. (2015). "Knowledge-based process management," in *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Gyor), 33–37.
- Tolgay, B., Dell'Orco, S., Maldonato, M. N., Vogel, C., Trojano, L., and Esposito, A. (2019). "EEGs as potential predictors of virtual agents' acceptance," in *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Naples), 433–438.
- Toma, M.-I., Rothkrantz, L. J., and Antonya, C. (2012). "Car driver skills assessment based on driving postures recognition," in *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)* (Kosice), 439–446.
- Török, Á., Török, Z. G., and Tölgyesi, B. (2017). "Cluttered centres: interaction between eccentricity and clutter in attracting visual attention of readers of a 16th century map," in *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Debrecen), 000433–000438.
- Török, M., Tóth, M. J., and Szöllosi, A. (2013). "Foundations and perspectives of mathability in relation to the CogInfoCom domain," in *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 869–872.
- Török, Z. G. and Török, Á. (2018). "Looking at the map - or navigating in a virtual city : interaction of visuospatial display and spatial strategies in VR," in *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 000327–000332.
- Török, Z. G. and Török, Á. (2019). "Remember the north - : Reference frames and spatial cognition at different scale," in *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Naples), 21–26.
- Vér, C. (2018). "3d VR spaces support R&D project management," in *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 000375–000378.
- Vicsi, K., Sztahó, D., and Kiss, G. (2012). "Examination of the sensitivity of acoustic-phonetic parameters of speech to depression," in *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)* (Kosice), 511–515.
- Vogel, C., and Esposito, A. (2019). "Linguistic and behavior interaction analysis within cognitive infocommunications," in *10th IEEE Conference on Cognitive Infocommunications* (Naples), 47–52.
- Vogel, C., and Esposito, A. (2020). Interaction analysis and cognitive infocommunications. *Infocommun. J.* 12, 2–9. doi: 10.36244/ICJ.2020.1.1
- Vogel, C., Hayes, E., Calawen, D., and Esposito, A. (2018). "Windfall scale, wealth consciousness and social proximity as influences on ultimatum game decisions," in *9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), eds P. Baranyi, A. Esposito, P. Földesi, and T. Mihálydeák, 185–190. doi: 10.1109/CogInfoCom.2018.8639893
- Wilcock, G., and Yamamoto, S. (2015). "Towards computer-assisted language learning with robots, Wikipedia and CogInfoCom," in *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Gyor), 115–119.



Automatic Estimation of Multidimensional Personality From a Single Sound-Symbolic Word

Maki Sakamoto^{1*}, Junji Watanabe² and Koichi Yamagata¹

¹ Department of Informatics, The University of Electro-Communications, Chofu, Japan, ² NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Atsugi, Japan

OPEN ACCESS

Edited by:

Carl Vogel,
Trinity College Dublin, Ireland

Reviewed by:

Tsutomu Fujinami,
Japan Advanced Institute of Science
and Technology, Japan
Maria Koutsombogera,
Trinity College Dublin, Ireland

*Correspondence:

Maki Sakamoto
maki.sakamoto@uec.ac.jp

Specialty section:

This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Psychology

Received: 18 August 2020

Accepted: 26 March 2021

Published: 22 April 2021

Citation:

Sakamoto M, Watanabe J and
Yamagata K (2021) Automatic
Estimation of Multidimensional
Personality From a Single
Sound-Symbolic Word.
Front. Psychol. 12:595986.
doi: 10.3389/fpsyg.2021.595986

Researchers typically use the “big five” traits (Extroversion, Agreeableness, Conscientiousness, Neuroticism, and Openness) as a standard way to describe personality. Evaluation of personality is generally conducted using self-report questionnaires that require participants to respond to a large number of test items. To minimize the burden on participants, this paper proposes an alternative method of estimating multidimensional personality traits from only a single word. We constructed a system that can convert a sound-symbolic word (SSW) that intuitively expresses personality traits into information expressed by 50 personality-related adjective pairs. This system can obtain information equivalent to the adjective scales using only a single word instead of asking many direct questions. To achieve this, we focused on SSWs in Japanese that have the association between linguistic sounds and meanings and express diverse and complex aspects of personality traits. We evaluated the prediction accuracy of the system and found that the multiple correlation coefficients for 48 personality-related adjective pairs exceeded 0.75, indicating that the model could explain more than half of the variations in the data. In addition, we conducted an evaluation experiment in which participants rated the appropriateness of the system output using a seven-point scale (with -3 as absolutely inappropriate and $+3$ as completely appropriate). The average score for 50 personality-related adjective pairs was 1.25. Thus, we believe that this system can contribute to the field of personality computing, particularly in terms of personality evaluation and communication.

Keywords: personality evaluation, system construction, sound-symbolic word, onomatopoeia, Japanese

INTRODUCTION

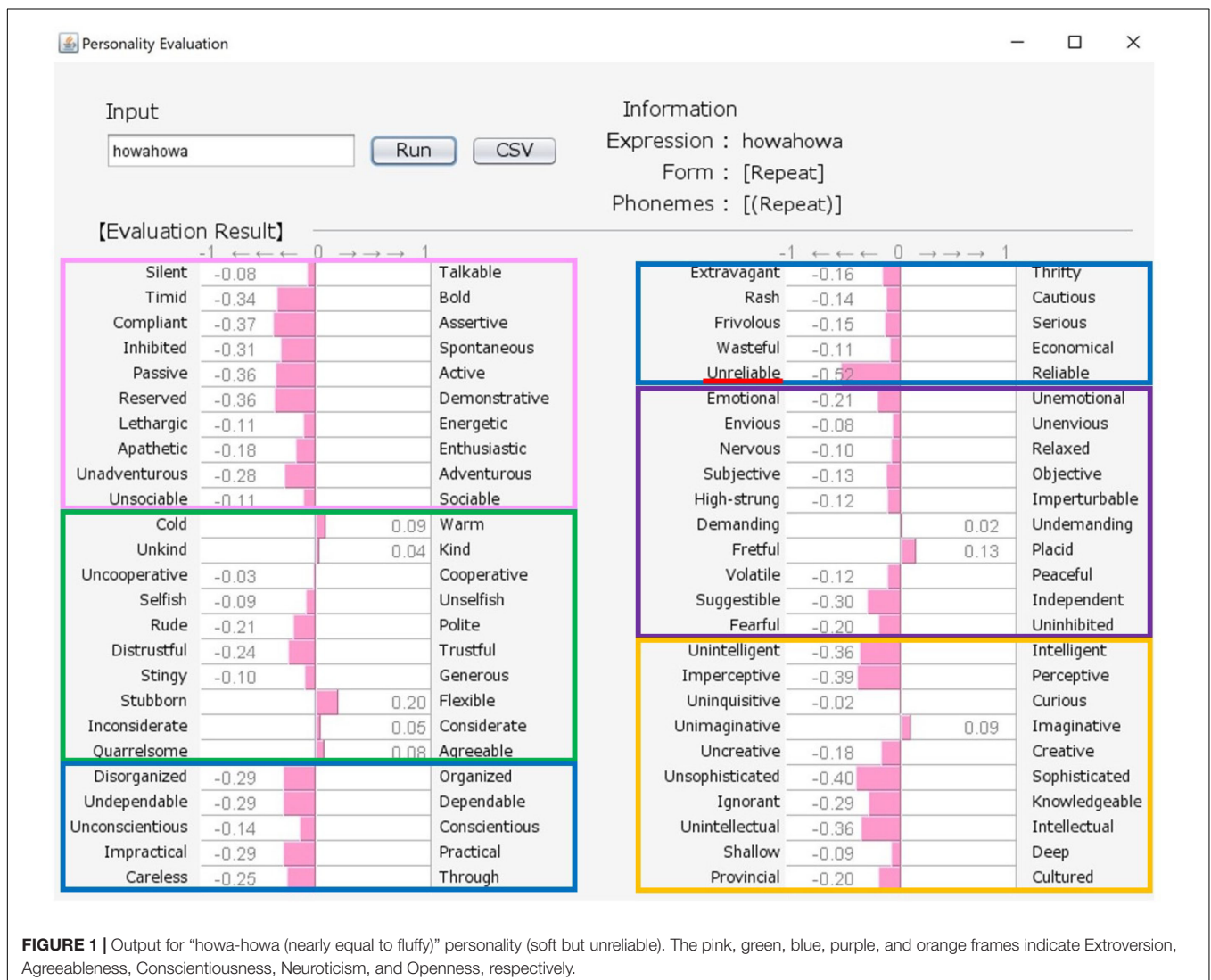
Understanding personality is crucial in making sense of our relationships with others. Most evaluations of personality traits, such as the “big five,” use questionnaires as self-report measures. However, these questionnaires generally ask people to respond to a large number of test items with two extreme positions or polar opposites described using adjectives. To ease the burden on participants, in this paper, we introduce a novel method for evaluating personality traits using only a single word. Specifically, we constructed a system that can convert a word that intuitively expresses personality traits into information equivalent to evaluations derived from 50 pairs of adjectives for big five personality traits. In other words, this system can generate the information of 50

adjectival scales from only a single word, instead of asking participants a lengthy series of questions. Our system is thus an efficient method for reducing the burden on people expressing their own personality, and therefore has applications to the field of personality computing that deal with personality evaluation and communication (see Vinciarelli and Mohammad, 2014 for a review).

The evaluation of the “big five” personality traits has a long history (Allport and Odbert, 1936; Cattell, 1943; Fiske, 1949; Tupes and Christal, 1961; Norman, 1967). In a landmark study, Goldberg (1982) extracted five factors; Extroversion, Agreeableness, Conscientiousness, Neuroticism, and Openness, and McCrae and Costa (1987) showed that these five factors could be reliably extracted regardless of whether the data were collected using self-rating or rating by others. The big five personality traits have since been evaluated using self-report questionnaires or the completion of questionnaires by others in studies conducted in a wide range of languages including English, German, Dutch, Czech, Polish, Russian, Italian, Spanish,

Hebrew, Hungarian, Turkish, Korean, Tagalog, and Japanese. Word-related personality studies are still frequently conducted worldwide. Since recent studies have identified systematic associations between personality and language use (Pennebaker and King, 1999; Hirsh and Peterson, 2009), there are a large-scale personality analyses using the standard word-based categories provided in the Linguistic Inquiry and Word Count (LIWC) 2001 program (Pennebaker et al., 2001). LIWC is the most commonly used language analysis program in studies investigating the relation between word use and psychological variables. Yarkoni (2010) analyzed personality and word use among bloggers focusing 66 LIWC categories. Das and Das (2016) developed a lexicon of words and phrases corresponding to each of the big five personality classes using LIWC text analysis tool (Mairesse et al., 2007), which categorizes words into psychologically meaningful categories, including personality-related groups.

Although there exist many reliable and widespread measures of personality, these are often questionnaires containing many adjectives, and can represent a burden on participants if they

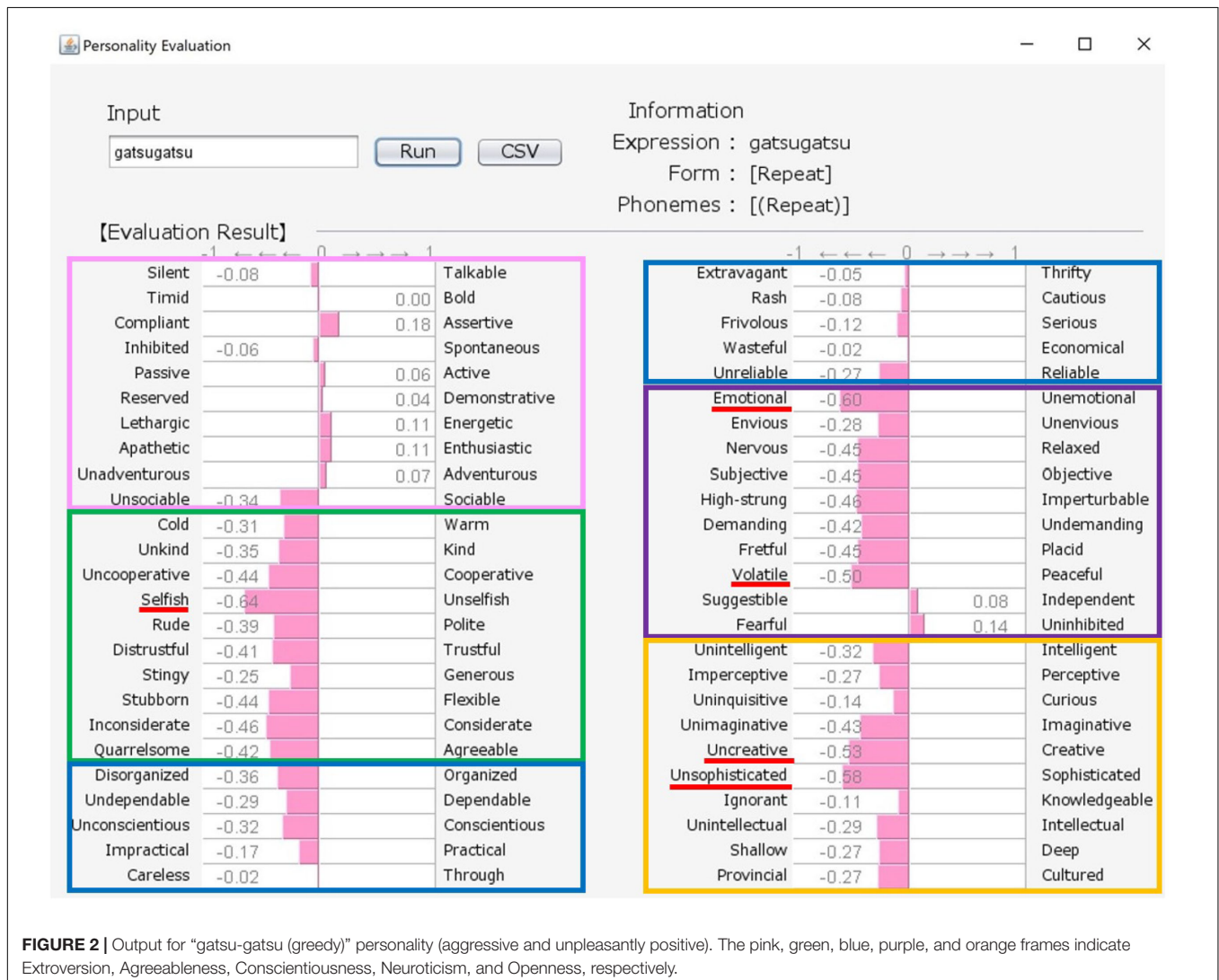


are asked to respond to a large number of test items. Here, we propose an engineering solution to this problem by building a new system that uses a word class referred to as “sound-symbolic words” (hereafter, SSWs). In the system, a SSW that can intuitively express personality traits is inputted, and evaluations against 50 personality-related adjective pairs are then produced based on analyses of the sounds of the inputted SSW (see samples of system output in **Figures 1, 2**). SSWs are adjective-like words that have associations between their sound and meaning. The existence of SSWs has been demonstrated in a wide variety of languages (e.g., Japanese, Chinese, Korean, Indonesian, Finnish, English, French, German, Modern Greek, and Native languages in North America, Latin America, Asia, Australia, and Africa).

There are two main reasons for our focus on SSWs. First, SSWs can describe a complex personality using a single word. For example, Japanese people use not only adjectives but also SSWs (called “onomatopoeia” in Japanese) that can describe complicated personality traits with a single word. “Honwaka” (a SSW in Japanese) means an agreeable, friendly,

and calm personality. A single SSW in Japanese tends to convey information that is more diverse and complex than what can be expressed by one adjective. Uchida (2005) reported that SSWs can indicate changes in impressions of personality that the big five cannot adequately describe, and Nishioka et al. (2006) suggested that SSWs can express aspects of personality that do not easily fit into the big five framework. There are many SSWs in Japanese. Komatsu et al. (2012) extracted about 120 personality-related SSWs from the Kojien word dictionary (5th edition), and extracted 60 SSWs in the format of a personality test.

Second, we focused on SSWs because they have systematic sound-symbolic features (Jespersen, 1922; Köhler, 1929; Sapir, 1929, for early studies; Hamano, 1998 for Japanese SSWs). For instance, the phonemes of Japanese SSWs, especially initial consonants, are known to characterize categories of touch. The two words “sara-sara” (smooth, dry, and comfortable) and “zara-zara” (rough, hard, and uncomfortable) differ only in their initial sounds (/s/ or /z/). This difference in only one feature can convey a critical



difference in the perception and evaluation of texture (e.g., Sakamoto and Watanabe, 2018). In a previous study (Doizaki et al., 2017), we used this sound-symbolic feature to evaluate complex sensations of touch, and developed a method for calculating the multidimensional ratings of a word by integrating the impressions of each phoneme. This system can convert a SSW in Japanese into quantitative ratings in multiple tactile dimensions. Another study also quantified the impressions of SSWs and used for description of robot motion (Ito et al., 2013).

In terms of information related to personality, sound-symbolic features can also be observed. The sounds of certain Japanese words are associated with features of personality traits (Shinohara and Kawahara, 2013). Lindauer (1990) demonstrated that English speakers judged the sound “takete” to be unfriendly and tough, whereas the sound “maluma” was considered to be friendly and tender. Milán et al. (2013) reported that Spanish speakers judged the word “kiki” to be clever and nervous. Kawahara et al. (2015) showed that Japanese and English speakers associate similar sound-symbolic features with personality traits: “difficult” personalities are typically associated with a phonetic class of sounds called obstruents. Thus, it appears that SSWs can be used as clues to evaluate and analyze personality traits based on their phonetic features. For these two reasons, we hypothesized that multidimensional ratings of personality could be predicted by combining the evaluations of each phoneme in a SSW, and that this could be advantageous for our system. In the next section, we

describe the construction and evaluation of the proposed system in detail.

SYSTEM CONSTRUCTION

We constructed a database that includes the sound-symbolic associations of phonemes and associated ratings of personality traits. In the experiment, participants viewed Japanese SSWs displayed on a monitor and rated their impressions of these words in terms of the 50 polar opposite adjective scales. From the results, we obtained a quantitative rating database for each phoneme of the SSWs, which enabled us to estimate impressions of a word by analyzing only the phonemes of the word. This database is the foundation for automatic conversion of Japanese SSWs into values related to the 50 pairs of adjectives of personality traits.

Participants

Thirty-two paid participants, aged 20–27 years (18 men and 14 women), participated in this experiment. They had no linguistics knowledge and were all native Japanese speakers. They had normal or corrected-to-normal vision, and none of them reported any visual or linguistic impairments. They were unaware of the purpose of the experiment, and informed consent was obtained from all the participants before the experiment started. All of the experiments, including those for system evaluation, were performed at the University of Electro-Communications,

TABLE 1 | All 126 sound-symbolic words (SSWs) used in the experiment (SSWs used in the study by Komatsu are shown with *).

1	ODO-ODO*	27	KEROQ*	53	SHIME-SHIME	79	GUNYAHRI	105	SABA-SABA*
2	SARARI*	28	FUNI-FUNI	54	OQTORI*	80	YUHRURI	106	SUI-SUI
3	HENA-HENA	29	HONWAKA*	55	UKI-UKI*	81	GIH-GIH	107	SARAQ*
4	NONBIRI*	30	CHARA-CHARA*	56	AWA-AWA	82	JIME-JIME*	108	KUNE-KUNE
5	DERE-DERE*	31	ZAWA-ZAWA	57	NUME-NUME	83	DARUN-DARUN	109	KII-KII
6	MOWA-MOWA	32	SHINA-SHINA	58	KIRIQ*	84	NIYAKERU*	110	HOWAHRI*
7	FUNWARI	33	KYAH-KYAH*	59	IRA-IRA*	85	SHIQKARI*	111	CHANTO*
8	SHEKA-SHEKA	34	WAKYA-WAKYA	60	NOHOHON*	86	YOWA-YOWA	112	BOKEQ*
9	SHAKIQ*	35	MOJI-MOJI*	61	DOSHIN	87	HONYA-HONYA	113	YORO-YORO
10	KUYO-KUYO*	36	KARAQ*	62	ORO-ORO*	88	TOGE-TOGE*	114	BUSUQ*
11	GONERU*	37	WAI-WAI	63	UKYA-UKYA	89	SUQKIRI*	115	KERORI*
12	FUHWAFUFWA	38	NUPU-NUPU	64	NECHIKOI*	90	DYURU-DYURU	116	GITYU-GITYU
13	KYAPI-KYAPI	39	GAMI-GAMI*	65	DAHRA-DAHRA	91	MUSUQ*	117	SAQPARI*
14	KARI-KARI*	40	GUHTARA	66	CHIMI-CHIM	92	BEQTARI*	118	FOWA-FOWA
15	FUHWARI	41	MOMA-MOMA	67	WAKI-WAKI	93	BONYARI*	119	ZUSHIN
16	CHIMA-CHIMA	42	WARA-WARA	68	BOUQ*	94	KICHIN*	120	NIKO-NIKO*
17	GASATSU*	43	KEBAI*	69	POKA-POKA	95	FUWA-FUWA*	121	SOWA-SOWA
18	POWA-POWA	44	FUNYA-FUNYA	70	RUN-RUN	96	YURU-YURU	122	CHARANPORAN*
19	KOSHO-KOSHO	45	YORE-YORE	71	SARAHRI	97	KAQ*	123	POWAN
20	YOTA-YOTA	46	SUPAQ	72	PEKO-PEKO	98	SUKAQ*	124	PERA-PERA*
21	WACHA-WACHA	47	PURA-PURA	73	ASHE-ASHE	99	KUTYI-KUTYI	125	MAQTARI*
22	PARIQ	48	GUFO-GUFO	74	BOSO-BOSO*	100	PIRI-PIRI*	126	KICHIQ*
23	FASA-FASA	49	MESO-MESO*	75	GATSUN	101	KUFO-KUFO		
24	NAYO-NAYO*	50	FUWAHN	76	RIN-RIN	102	FUWAN		
25	BISHIQ*	51	UJI-UJI*	77	KIQCHIRI*	103	AQSARI*		
26	BIKU-BIKU*	52	BETA-BETA*	78	RAN-RAN	104	SHIO-SHIO		

TABLE 2 | Fifty rating scales for evaluating personality, with examples of values for “howa”.

Big Five	Rating scales	Consonants		Vowels		R	R _{SSW}
		/h/ (X ₁)	/w/ (X ₇)	/o/ (X ₄)	/a/ (X ₁₀)		
Extroversion	Silent – Talkable	−0.074	0.025	−0.156	0.114	0.601	0.807
	Timid – Bold	−0.048	−0.231	−0.100	0.073	0.622	0.814
	Compliant – Assertive	−0.063	−0.237	−0.096	0.011	0.617	0.823
	Inhibited – Spontaneous	−0.112	−0.080	−0.135	0.029	0.643	0.822
	Passive – Active	−0.132	−0.119	−0.154	0.050	0.644	0.818
	Reserved – Demonstrative	−0.096	−0.145	−0.147	0.074	0.622	0.839
	Lethargic – Energetic	−0.108	0.016	−0.104	0.040	0.616	0.812
	Apathetic – Enthusiastic	−0.120	0.084	−0.126	0.000	0.523	0.767
	Unadventurous – Adventurous	−0.050	−0.192	−0.107	0.084	0.596	0.804
	Unsociable – Sociable	0.037	−0.018	−0.115	−0.022	0.563	0.778
Agreeableness	Cold – Warm	0.029	0.050	−0.005	0.012	0.527	0.778
	Unkind – Kind	0.046	0.021	−0.003	−0.013	0.541	0.779
	Uncooperative – Cooperative	0.044	0.021	−0.018	−0.012	0.544	0.767
	Selfish – Unselfish	0.061	0.046	0.023	0.012	0.495	0.791
	Rude – Polite	−0.007	0.030	0.001	−0.090	0.610	0.826
	Distrustful – Trustful	0.017	−0.002	−0.023	−0.068	0.584	0.793
	Stingy – Generous	0.013	−0.147	−0.031	0.064	0.519	0.778
	Stubborn – Flexible	0.069	0.019	−0.040	0.040	0.539	0.800
	Inconsiderate – Considerate	0.072	0.040	−0.005	−0.048	0.545	0.801
	Quarrelsome – Agreeable	0.027	−0.019	0.033	0.004	0.580	0.806
Conscientiousness	Disorganized – Organized	0.075	−0.134	−0.047	−0.062	0.528	0.826
	Undependable – Dependable	−0.017	−0.059	−0.019	−0.087	0.602	0.796
	Unconscientious – Conscientious	0.033	−0.039	0.021	−0.060	0.582	0.804
	Impractical – Practical	0.015	−0.065	−0.044	−0.064	0.562	0.805
	Careless – Through	−0.042	0.036	−0.020	−0.106	0.499	0.780
	Extravagant – Thrifty	0.017	0.038	−0.010	−0.125	0.505	0.812
	Rash – Cautious	−0.057	0.153	0.010	−0.138	0.494	0.827
	Frivolous – Serious	0.007	0.013	0.030	−0.141	0.579	0.817
	Wasteful – Economical	0.061	0.025	0.017	−0.113	0.509	0.824
	Unreliable – Reliable	−0.061	−0.181	−0.028	−0.034	0.608	0.792
Neuroticism	Emotional – Unemotional	0.061	−0.210	0.084	−0.001	0.562	0.819
	Envious – Unenvious	0.000	−0.214	0.068	0.017	0.579	0.810
	Nervous – Relaxed	0.019	−0.150	0.031	0.072	0.582	0.776
	Subjective – Objective	0.021	−0.076	0.033	0.018	0.500	0.822
	High-strung – Imperturbable	0.023	−0.240	0.045	0.056	0.550	0.789
	Demanding – Undemanding	0.019	−0.016	0.057	0.008	0.553	0.802
	Fretful – Placid	0.049	−0.045	0.076	0.013	0.593	0.799
	Volatile – Peaceful	0.127	−0.165	0.059	−0.037	0.536	0.825
	Suggestible – Independent	−0.020	−0.256	−0.110	0.026	0.542	0.816
	Fearful – Uninhibited	−0.047	−0.255	−0.103	0.114	0.537	0.795
Openness	Unintelligent – Intelligent	−0.045	−0.162	−0.074	−0.011	0.539	0.784
	Imperceptive – Perceptive	0.003	−0.126	−0.163	−0.038	0.555	0.811
	Uninquisitive – Curious	−0.074	0.065	−0.065	−0.014	0.444	0.736
	Unimaginative – Imaginative	0.008	0.095	0.030	−0.019	0.435	0.718
	Uncreative – Creative	−0.038	0.049	−0.026	−0.017	0.442	0.784
	Unsophisticated – Sophisticated	0.000	−0.014	−0.084	−0.088	0.597	0.839
	Ignorant – Knowledgeable	−0.010	0.019	−0.044	−0.084	0.496	0.775
	Unintellectual – Intellectual	−0.008	0.015	−0.018	−0.096	0.549	0.806
	Shallow – Deep	0.005	0.128	0.055	−0.085	0.452	0.793
	Provincial – Cultured	−0.023	0.037	−0.013	−0.067	0.526	0.770

Japan. Experimental procedures were conducted in accordance with the ethical standards outlined by the Declaration of Helsinki.

Stimuli

To obtain sound-symbolic associations of all Japanese phonemes with the 50 pairs of adjectives of the big five, we selected word stimuli that included all varieties of Japanese phonemes. First, we decided to use all 60 SSWs given in Komatsu et al. (2012). Then, we added an additional 66 SSWs to cover all Japanese phonemes. The total number of SSWs used in the experiment was 126, as shown in **Table 1**. Each SSW was evaluated by ten participants. That is, the 126 SSWs were divided into 5 groups (25 or 26 words each), and 18 of the 32 participants evaluated 2 different groups, while 14 participants evaluated 1 group.

The participant sat in front of a 23-inch LCD monitor. The SSWs were presented on the monitor with a resolution of 1024×768 pixels and a refresh rate of 60 Hz. The SSWs were presented in 11-pt MS PGothic font. The distance between the participant's eyes and the screen was approximately 50 cm. The

polar opposite scales were presented in the form of an answer matrix on the monitor. The rating scales, shown in **Table 2**, included 50 pairs of adjectives used for expressing personality traits of the big five. Each participant responded to all 50 rating scales for each SSW.

Procedure

The trials started with the presentation of a SSW on the monitor. The participants were asked to report how they felt about each word on a seven-point SD scale, e.g., for the silent-talkable scale, participants selected one of the following seven points: -3, very silent; -2, silent; -1, slightly silent; 0, neither; +1, slightly talkable; +2, talkable; and, +3, very talkable. The participants responded by pushing one of seven buttons. The time allotted for answering was unlimited, but most participants took less than 1 min per trial. The presentation order of the SSWs was randomized among participants. The order and polarity of the scales were also randomized in the answer matrix.

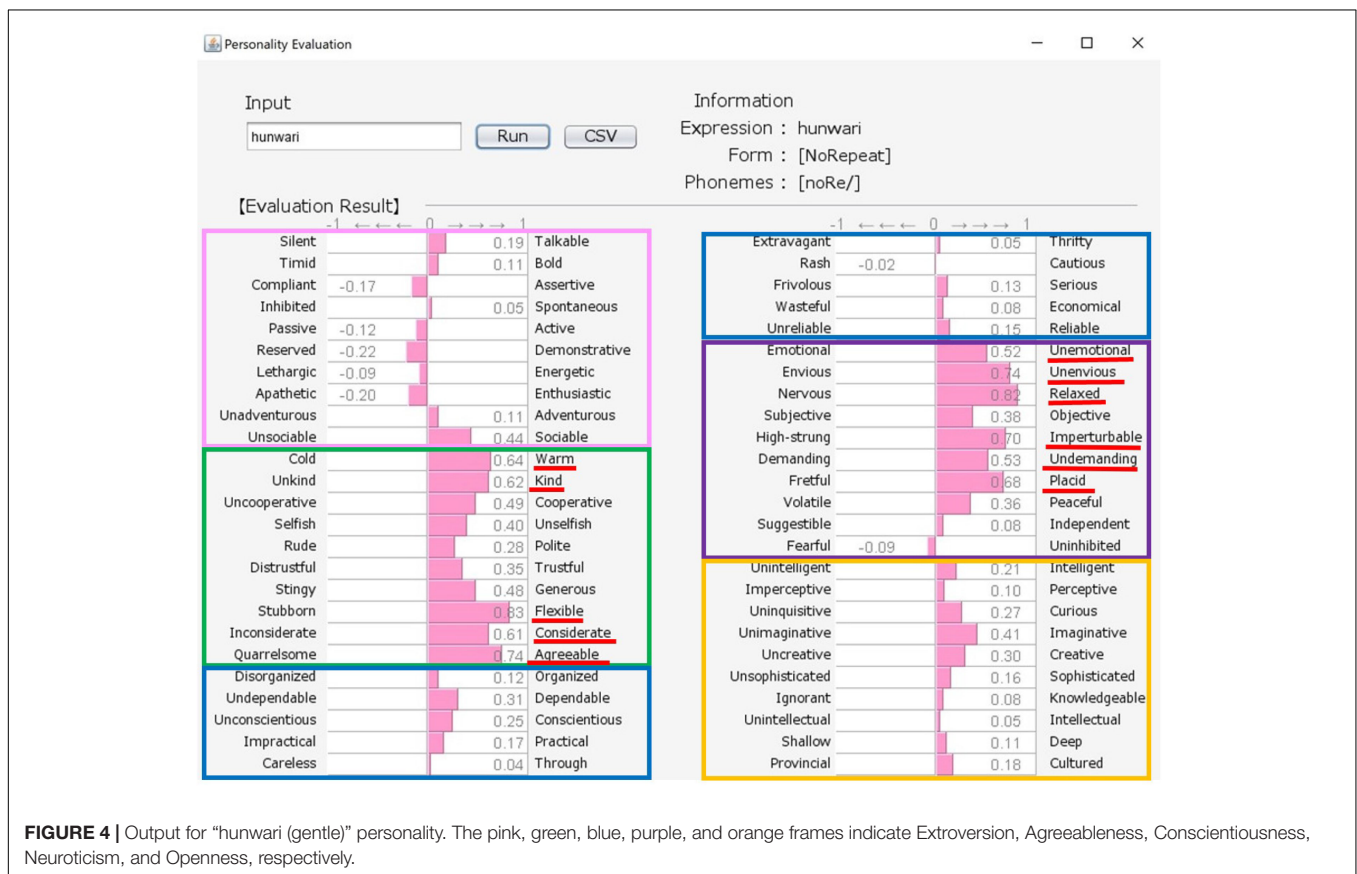
TABLE 3 | Correspondence between variables and phonemes.

First syllable	Second syllable	Phonological characteristics	Variation of phonemes
X_1	X_7	Consonants	/k/, /s/, /t/, /n/, /h/, /m/, /y/, /r/, /w/ or absence
X_2	X_8	Voiced sounds, /p/-sounds	Presence (/g/, /z/, /d/, /b/, /p/) or absence
X_3	X_9	Contracted sounds	Presence(/ky/, /sy/, /ty/, /ny/, /hy/, /my/, /ry/, /gy/, /zy/, /by/, /py/) or absence
X_4	X_{10}	Vowels	/a/, /i/, /u/, /e/, /o/
X_5	X_{11}	Semi-vowels	/a/, /i/, /u/, /e/, /o/ or absence
X_6	X_{12}	Special sounds	/N/, /Q/, /R/, /L/ or absence
X_{13}		Repetition	Presence(ex. huwa-huwa) or absence

SSW described for person 1	Rating scales	System output values	Appropriateness of system output						
			-3	-2	-1	0	1	2	3
howahowa	Silent – Talkable	-0.08							
howahowa	Timid – Bold	-0.34							
howahowa	Compliant – Assertive	-0.37							
howahowa	Inhibited – Spontaneous	-0.31							
howahowa	Passive – Active	-0.36							
howahowa	Reserved – Demonstrative	-0.36							
howahowa	Lethargic – Energetic	-0.11							
howahowa	Apathetic – Enthusiastic	-0.18							
howahowa	Unadventurous – Adventurous	-0.28							
howahowa	Unsociable – Sociable	-0.11							
howahowa	Cold – Warm	0.09							
howahowa	Unkind – Kind	0.04							
howahowa	Uncooperative – Cooperative	-0.03							
howahowa	Selfish – Unselfish	-0.09							
howahowa	Rude – Polite	-0.21							
howahowa	Distrustful – Trustful	-0.24							
howahowa	Stingy – Generous	-0.1							
howahowa	Stubborn – Flexible	0.2							
howahowa	Inconsiderate – Considerate	0.05							
howahowa	Quarrelsome – Agreeable	0.08							
howahowa	Disorganized – Organized	-0.29							
howahowa	Undependable – Dependable	-0.29							
howahowa	Unconscientious – Conscientious	-0.14							
howahowa	Impractical – Practical	-0.29							
howahowa	Careless – Thorough	-0.25							
howahowa	Extravagant – Thrifty	-0.16							
howahowa	Rash – Cautious	-0.14							

FIGURE 3 | Example of evaluation form (SSW described for person X; rating scales; system output value; and appropriateness of system output).

Big Five	Rating scales	Mean	Z	Big Five	Rating scales	Mean	Z
Extroversion	Silent – Talkable	1.07	3.77	Conscientiousness	Extravagant – Thrifty	1.17	4.17
	Timid – Bold	0.90	3.31		Rash – Cautious	1.14	4.35
	Compliant – Assertive	1.52	6.57		Frivolous – Serious	1.24	4.62
	Inhibited – Spontaneous	1.21	4.72		Wasteful – Economical	1.00	3.70
	Passive – Active	1.55	6.44	Neuroticism	Unreliable – Reliable	0.88	2.92
	Reserved – Demonstrative	1.40	6.13		Emotional – Unemotional	1.24	5.29
	Lethargic – Energetic	1.14	4.89		Envious – Unenvious	1.71	7.64
	Apathetic – Enthusiastic	0.98	3.74		Nervous – Relaxed	1.83	9.80
	Unadventurous – Adventurous	0.69	2.78		Subjective – Objective	1.60	8.36
Agreeableness	Unsociable – Sociable	1.38	5.70	High-strung – Imperturbable	1.40	6.05	
	Cold – Warm	1.50	5.82	Demanding – Undemanding	1.24	5.43	
	Unkind – Kind	1.26	4.61	Fretful – Placid	1.55	6.57	
	Uncooperative – Cooperative	1.38	5.17	Volatile – Peaceful	1.79	9.21	
	Selfish – Unselfish	1.12	4.59	Suggestible – Independent	1.67	9.09	
	Rude – Polite	1.26	3.98	Openness	Fearful – Uninhibited	1.29	5.36
	Distrustful – Trustful	1.10	3.81		Unintelligent – Intelligent	1.57	6.13
	Stingy – Generous	1.29	5.29		Imperceptive – Perceptive	1.64	6.90
	Stubborn – Flexible	1.69	8.30		Uninquisitive – Curious	1.26	5.04
Inconsiderate – Considerate	1.31	4.96	Unimaginative – Imaginative		0.93	4.05	
Quarrelsome – Agreeable	1.50	6.04	Uncreative – Creative		0.67	2.66	
Conscientiousness	Disorganized – Organized	1.81	11.61		Unsophisticated – Sophisticated	0.69	2.73
	Undependable – Dependable	0.88	2.72	Ignorant – Knowledgeable	0.83	3.36	
	Unconscientious – Conscientious	1.12	3.72	Unintellectual – Intellectual	0.64	2.33	
	Impractical – Practical	1.19	4.62	Shallow – Deep	0.81	3.50	
	Careless – Through	1.40	5.78	Provincial – Cultured	1.17	4.86	



Personality Estimation Model

The experimental results produced 63,000 datapoints (50 rating scales \times 126 words \times 10 participants). To estimate the impressions of SSWs, we created a linear regression model in which the following equation was used to predict each rating value:

$$Y = \sum_{i=1}^{13} X_i + \text{Const.} \quad (1)$$

where Y represents the rating values of the respective 50 scales, and $X_1 - X_{13}$ are quantified values of phonemes. $X_1 - X_6$, respectively are the values of the specific consonant, voiced sound/p-sound, contracted sounds, vowels, semivowels, and special phonemes in the first syllable. $X_7 - X_{12}$ represent the same categories for the second syllable, respectively, and X_{13} denotes the presence or absence of repetitions in the word (see also **Table 3**). Using the rating values as the objective variables and the variation of phonemes as the predictor variables, we conducted mathematical quantification theory class I, which is a type of multiple regression analysis.

Table 2 shows examples of the results of the SSW “howa.” According to Equation (1), the rating values of a SSW can be calculated by the linear sum of the values ($X_1 - X_{13}$) of the

word. The expression “howa” is composed of the first mora /ho/ (/h/ + /o/) and the second mora /wa/ (/w/ + /a/). Therefore, the value of the unreliable-reliable scale on a seven-point scale (unreliable -3 to reliable 3) divided by three is estimated by the following equation [see Equation (1) and **Table 2**]. The estimated value of -0.304 suggests that “howa” is associated with an unreliable personality.

$$Y = /h/ + /o/ + /w/ + /a/ + \text{Cosnt.}$$

$$\begin{aligned} &= /h/ (X_1) + \text{absence} (X_2) + \text{absence} (X_3) + /o/ (X_4) \\ &\quad + \text{absence} (X_5) + \text{absence} (X_6) + /w/ (X_7) + \text{absence} (X_8) \\ &\quad + \text{absence} (X_9) + /a/ (X_{10}) + \text{absence} (X_{11}) + \text{absence} (X_{12}) \\ &\quad + \text{absence} (X_{13}) + \text{Const.} \\ &= (-0.061) + (0.079) + (0.02) + (-0.028) + (-0.005) \\ &\quad + (-0.042) + (-0.181) + (0.035) + (0.007) + (-0.034) \\ &\quad + (0.013) + (-0.033) + (0.133) + (-0.207) \\ &= -0.304 \end{aligned}$$

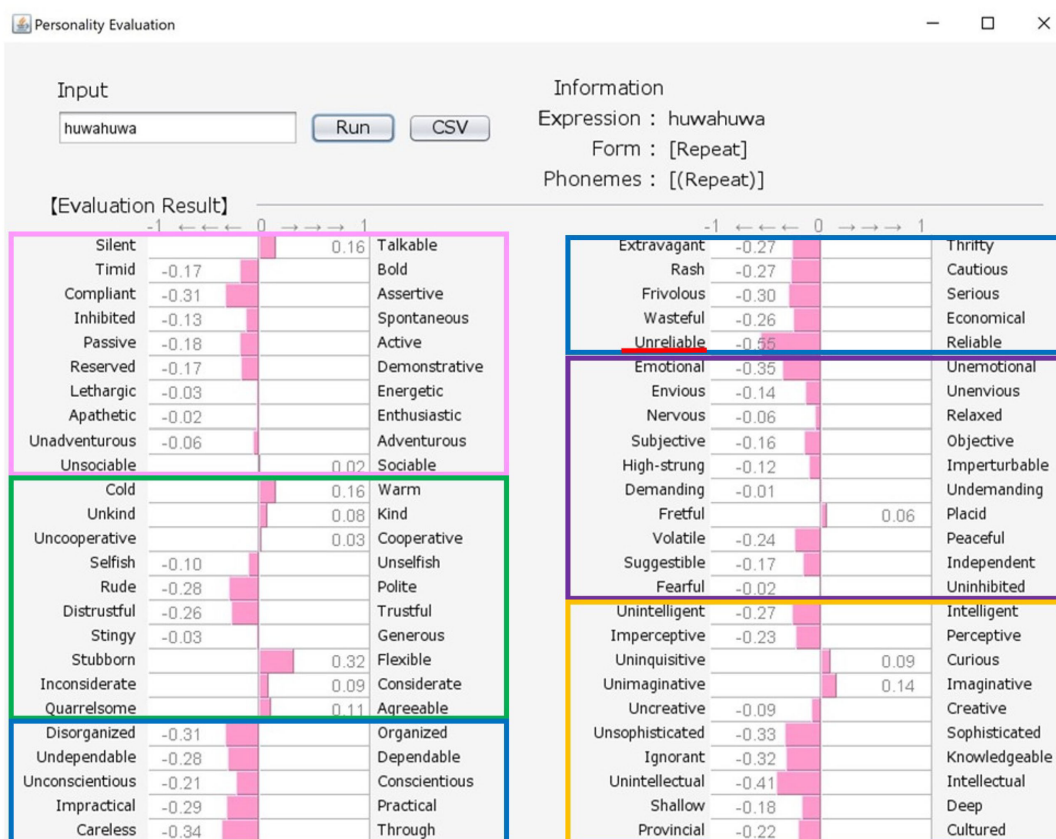


FIGURE 5 | Output for “huwahuwa (nearly equal to fluffy)” personality. The pink, green, blue, purple, and orange frames indicate Extroversion, Agreeableness, Conscientiousness, Neuroticism, and Openness, respectively.

The multiple correlation coefficient, R, shown in **Table 2** is calculated by

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{RSS}{ESS + RSS}, \quad (2)$$

where RSS is the residual sum of squares, and TSS is the total sum of squares that can be partitioned into the explained sum of squares (ESS) and the RSS, i.e.,

$$TSS = ESS + RSS, \quad (3)$$

due to the Pythagorean equation in the sample space \mathbb{R}^n with $n = 1260$. Because ten participants in the experiment evaluated each SSW, each RSS could be further partitioned into the sum of squares for the residual of participant differences, RSS_p , and the residual of SSW differences, RSS_{SSW} , i.e.,

$$RSS = RSS_p + RSS_{SSW}. \quad (4)$$

Note that averaging the ratings of the 10 participants can be regarded as an orthogonal projection in the sample space \mathbb{R}^n ,

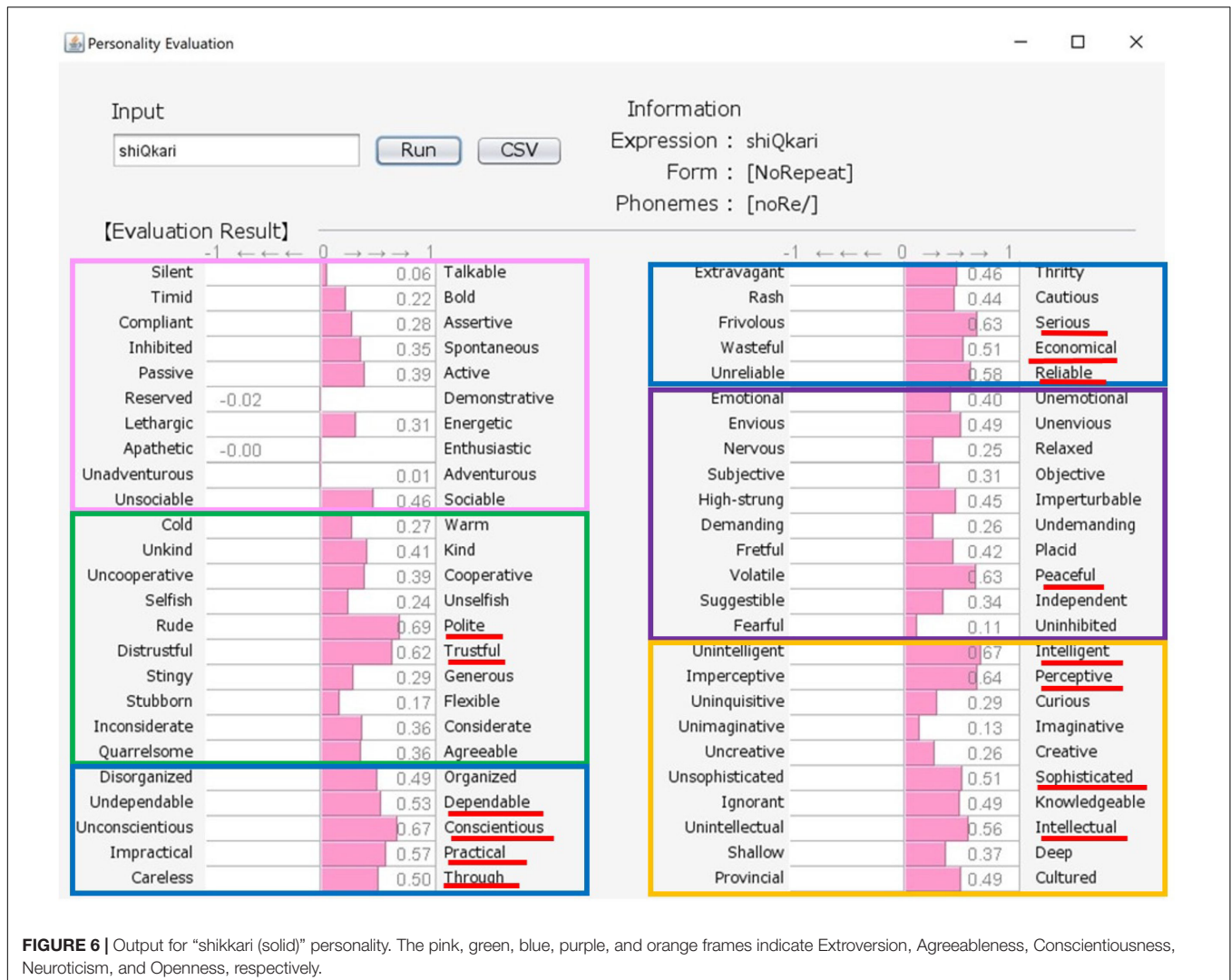
as in the linear regression. Thus, the Pythagorean Equations (3) and (4) are possible. The multiple correlation coefficients without participant differences, R_{SSW} , can be calculated by

$$R_{SSW}^2 = 1 - \frac{RSS_{SSW}}{ESS + RSS_{SSW}}. \quad (5)$$

The multiple correlation coefficients, R_{SSW} , were used as indicators of prediction accuracy. We can see that all ten multiple values of R_{SSW} for Extroversion, Agreeableness, Conscientiousness, Neuroticism, and 8 of the 10 for Openness exceeded 0.75, indicating that the model explains more than half of variations in the data.

User Interface and Information Processing

Our system comprises a user interface module, a SSW-parsing module, an analyzing module, and a database. **Figures 1, 2** show the system's estimation values for “howa-howa (nearly equal to fluffy)” and “gatsu-gatsu (greedy).” When a user inputs a SSW



into the text field in the upper-left frame of a window and presses the “Run” button, the parsing module automatically divides the word into each phoneme and classifies its form. On the basis of (1) and the database, the analyzing module calculates the rating values of the word for the 50 scales. Then, the module converts the calculated values from -3 to 3 into values from -1 to 1 . Finally, graphs of the estimated values of the word are displayed in the lower frame. The form and phonemic elements of the word are displayed in the upper-right frame.

SYSTEM EVALUATION

Here we describe an experiment we conducted to verify the validity of the system constructed in this study.

Participants

Seven participants aged 22–27 years (4 men and 3 women) participated in the evaluation experiment. The participants

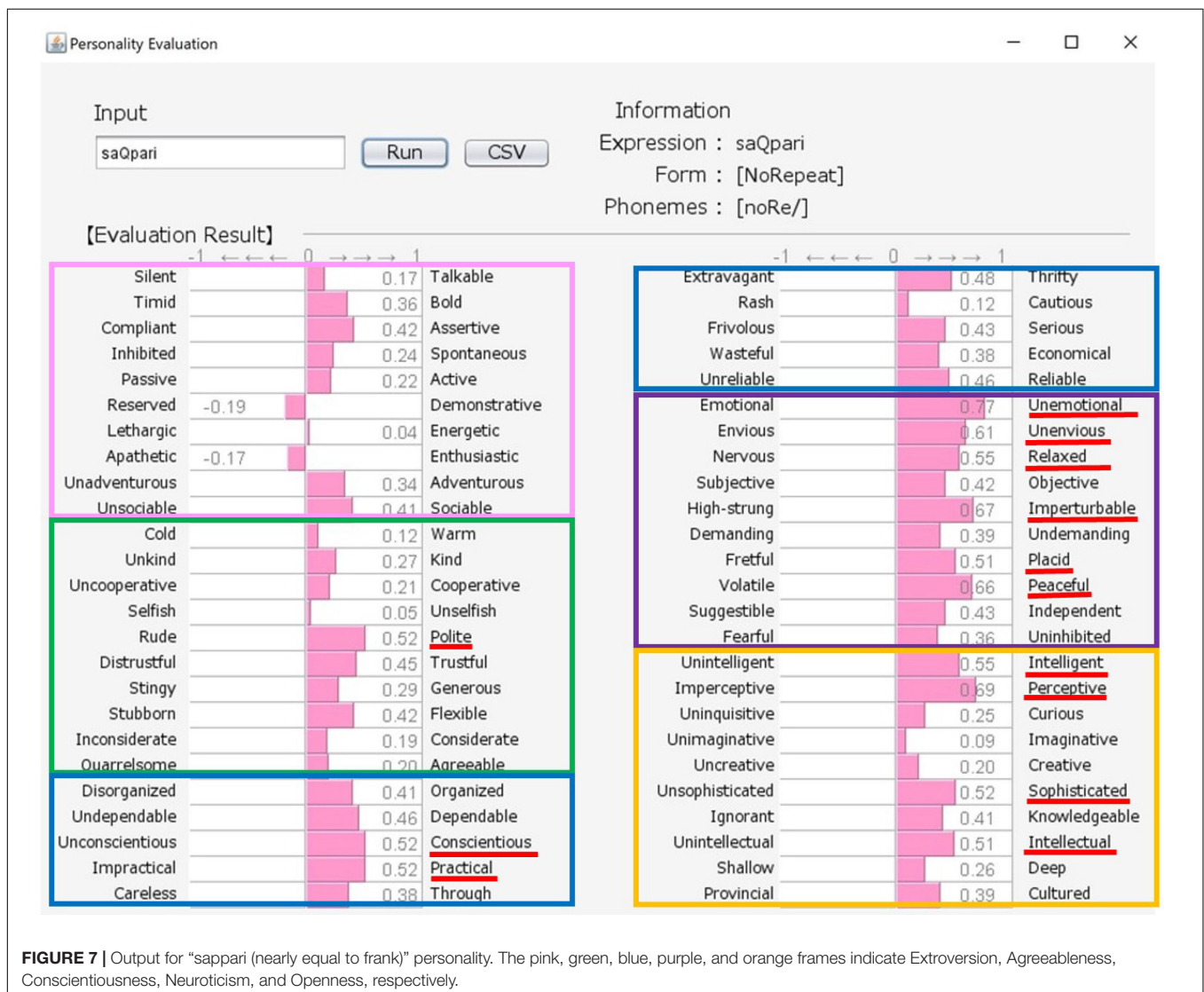
were familiar with each other so that each participant could evaluate the personality of the target person on the basis of their friendship.

Procedure

Each of the seven participants spontaneously described the personality traits of the other six people using SSWs. Thus, a total of 42 SSWs were obtained. The 42 SSWs were then analyzed by the system constructed in this study, and values of the 50 scales were obtained. Each participant also rated the output values of the 50 scales for each of the other six persons from -3 as “absolutely inappropriate” to $+3$ as “completely appropriate.” **Figure 3** shows an example of the evaluation form.

Results

In this evaluation experiment, each scale was evaluated 42 times, which is sufficient for the central limit theorem to be applied. Therefore, the average scores followed normal

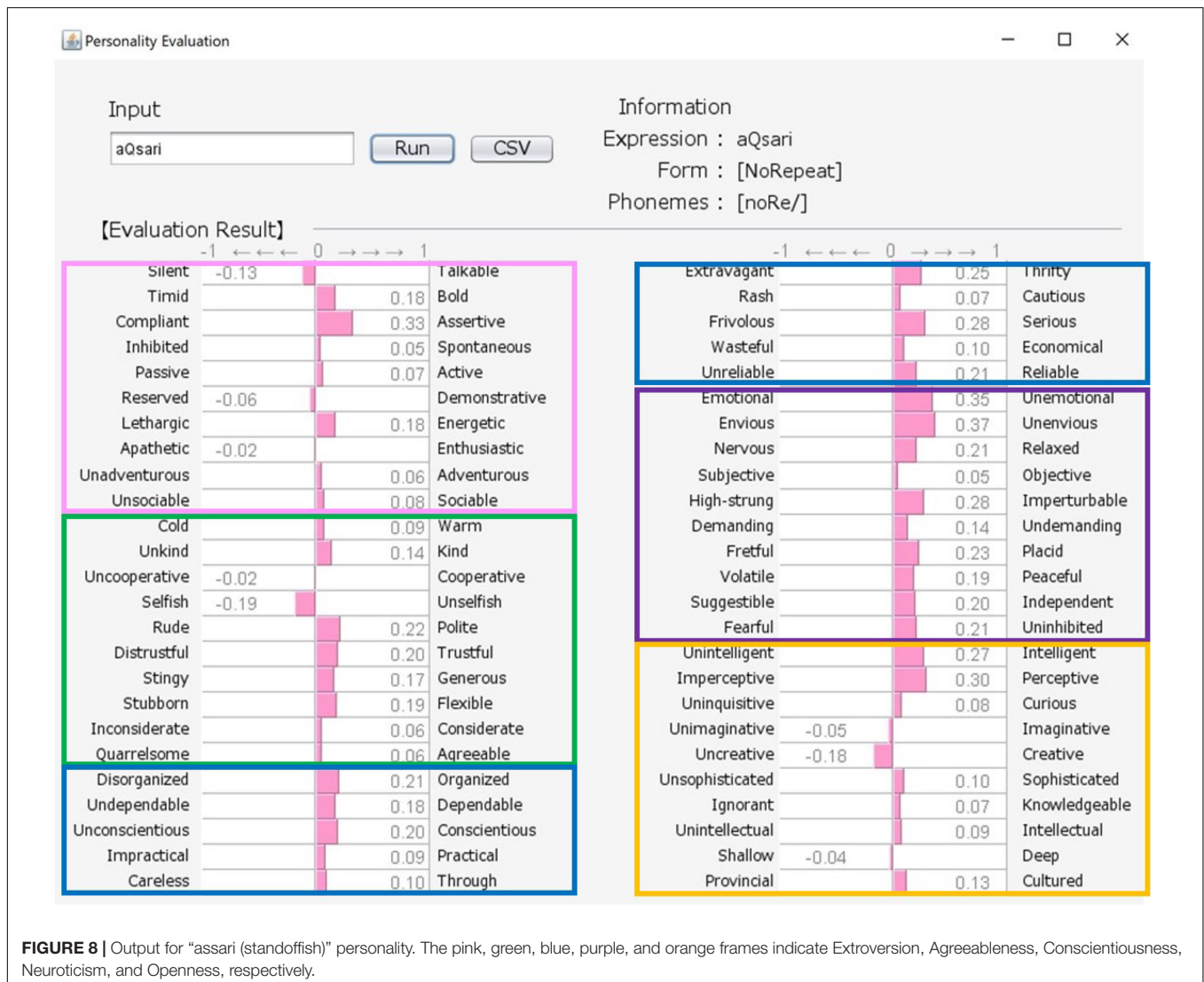


distributions, and the Z-test was valid. **Table 4** shows the average scores and Z-values of the Z-tests for each scale. Because the Z-values for all 50 scales exceeded 2, the average scores were significantly above zero at the 0.025 level, indicating that the participants tended to give positive evaluations to all the scales. The average scores for Extroversion, Agreeableness, Conscientiousness, Neuroticism, and Openness were 1.19, 1.34, 1.18, 1.53, and 1.02, respectively. The average Z-values for Extroversion, Agreeableness, Conscientiousness, Neuroticism, and Openness were 4.80, 5.26, 4.82, 7.28, and 4.16, respectively. Although the scores for Openness were slightly lower than the others, this tendency was similar to that observed for the multiple correlation coefficient (R and R_{SSW} in **Table 2**).

DISCUSSION

Many studies of personality have been based on the “big five” hypothesis proposed by Goldberg (1982), which suggests

that human personality is classified into five factors derived from 100 adjectives. A questionnaire method for evaluating personality using these five categories is well-established and widely used today. Costa and MacCrae (1992); Goldberg (1992)Goldberg (1993), and Goldberg et al. (2006) further developed the scale to measure personality based on big five personality traits. However, personality evaluations using lengthy questionnaires can burden the respondents. There are many online sites based on big five theory, which usually take around 10 min. The following sites are examples: <https://openpsychometrics.org/tests/IPIP-BFFM> accessed March 12, 2021, which is referring to Goldberg (1992), and <https://www.123test.com/personality-test/>, accessed February 28, 2021. In contrast, in our system, when a word that intuitively expresses personality is inputted into the text field, information equivalent to evaluations against multiple personality-related adjectives is instantly generated on the basis of analyses of the sounds in the word. All the participant has to do is provide only one word for an answer. We

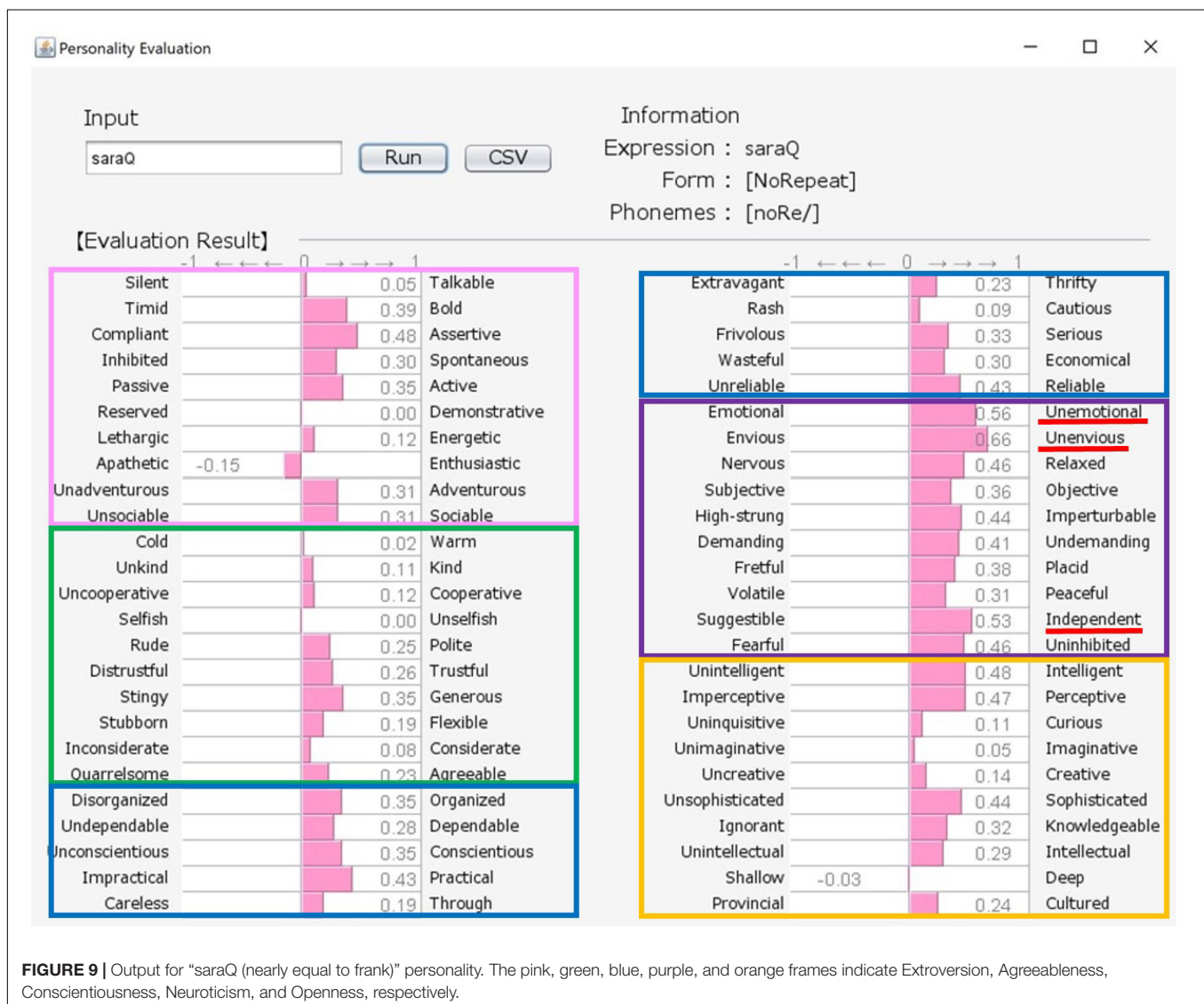


believe that our system represents an alternative method for evaluating personality.

The proposed system might function effectively not only for self-evaluation but also in situations where personality evaluation by others is required. For instance, it could be particularly valuable when one needs to understand someone's personality with limited available information. For example, in a business situation, a salesperson may understand elements of a customer's personality after only a brief description given by colleagues. In an educational setting, a new teacher could understand a student's personality through descriptions given by other students. Our study expands personality research into the field of engineering application, and proposes a novel system for capturing complex personality traits.

Japanese people frequently use SSWs to express personality, such as “howa-howa,” which indicates a soft but unreliable person. SSWs are used in everyday conversation and a single SSW tends to convey more information than one adjective,

as shown in the system output example for “howa-howa” in **Figure 1**. Japanese has a large SSW vocabulary that can express complex details about personality. At the same time, meanings of SSWs are typically characterized by phonetic features associated with multiple sensory experiences (e.g., Sakamoto and Watanabe, 2016 for taste; Sakamoto and Watanabe, 2018 for touch). In addition, Kawahara et al. (2015) showed a direct relationship between sensory impression (visual shape) and personality. We tested the relationship between tactile impression and personality using a Google search. For 60 SSWs given in Komatsu et al. (2012), we used the search terms “SSW person” (for personality expression) and “SSW touch” (for tactile expression) in a Google search query on May 7, 2020. More than 10,000 search results were obtained for both personality and tactile texture for the following 8 SSWs: “hunwari” and “huwa-huwa” (SSWs related to softness); “shikkari” (related to hardness); and “sappari,” “assari,” “saraQ,” “sarari,” and “sukkiri” (related to dryness). Although these



expressions are all used for both tactile aspects and personality, they behave differently. “Hunwari” and “huwa-huwa” are both related to softness of touch. However, as shown in **Figures 4, 5**, the “hunwari” personality (nearly equal to gentle personality) is warm, kind, flexible, and relaxed, while the “huwa-huwa” personality (nearly equal to fluffy personality) is unreliable. The “hunwari” personality seems more positive. The meanings of “shikkari” in terms of touch and personality are likely to share commonalities. “Shikkari” is related to hardness of touch, while the “shikari” personality (solid personality) is, as shown in **Figure 6**, polite, trustful, conscientious, serious, and intelligent. “Sappari,” “assari,” “saraQ,” “sarari,” and “sukkiri” are all related to dryness of touch, while their meanings for personality are different but nearly equal to frank, as shown in **Figures 7–11**.

A limitation of our system is that it is available only in Japanese, because the application was constructed using Japanese SSWs. However, it has been argued that sensory-sound

associations for auditory, visual, tactile, and olfactory perceptions are universal phenomena observed in many languages, especially Asian-African languages (Dingemanse, 2012). Sound-symbolic features also occur in Indo-European languages, such as English (Bloomfield, 1933; Bolinger, 1950; Crystal, 1995). For example, roughly half of the English words starting with “gl-” (such as, glance, glare, glass, glimpse, and glow) imply something visual and bright (Crystal, 1995). Studies have found world-wide sound symbolism in words referring to visual shapes, such as “mal” vs. “mil” (Sapir, 1929) and “bouba” vs. “kiki” (Ramachandran and Hubbard, 2001) for round vs. sharp shapes.

We believe that sensory-sound associations for personality description have been overlooked, although there are languages that use SSWs to express personality. For example, the SSW “lumbud-lumbud” in Mundari (the South Asian linguistic area), whose original meaning is the appearance of a hole opening and closing, can also refer to a person that cannot keep a

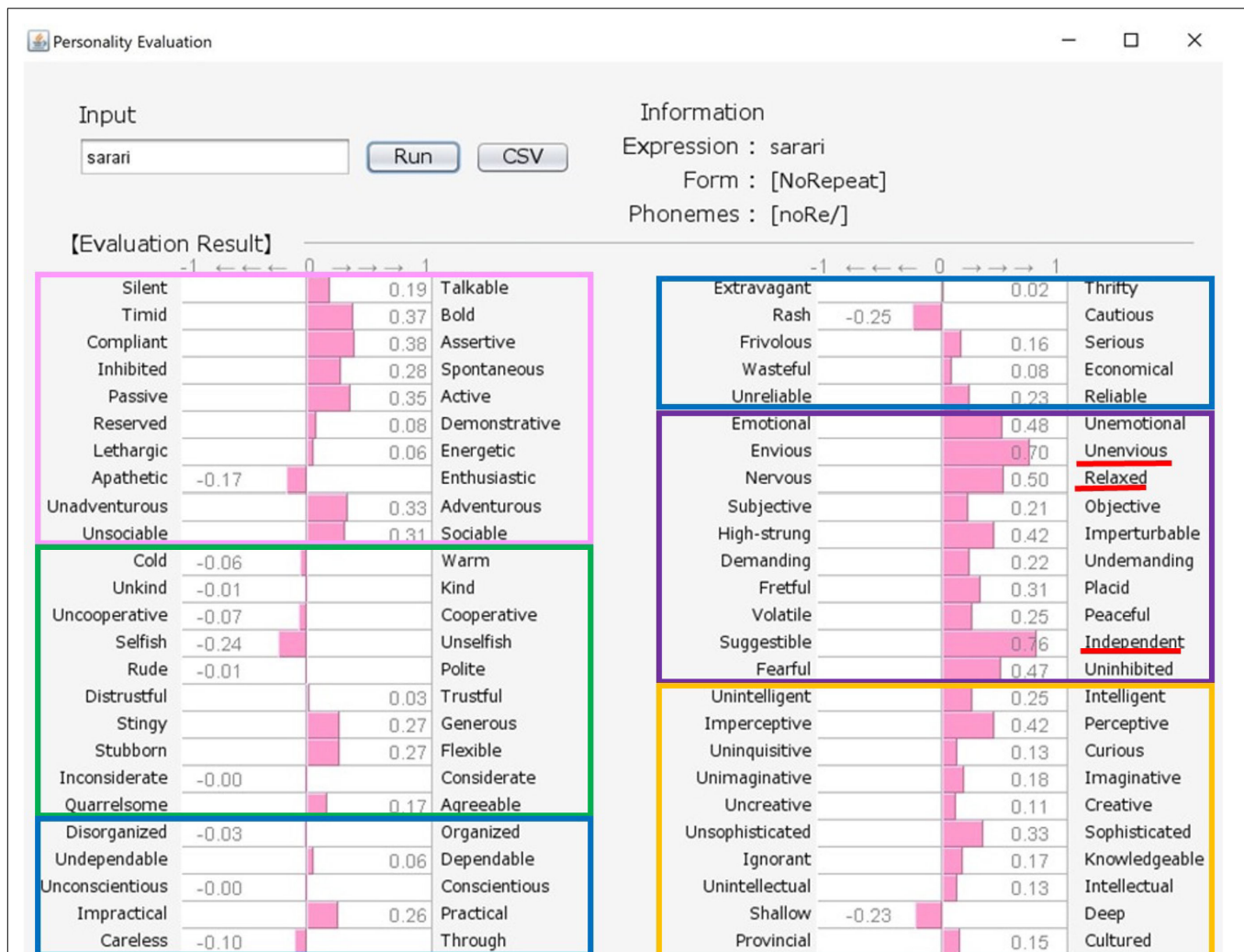


FIGURE 10 | Output for “sarari (nearly equal to frank)” personality. The pink, green, blue, purple, and orange frames indicate Extroversion, Agreeableness, Conscientiousness, Neuroticism, and Openness, respectively.

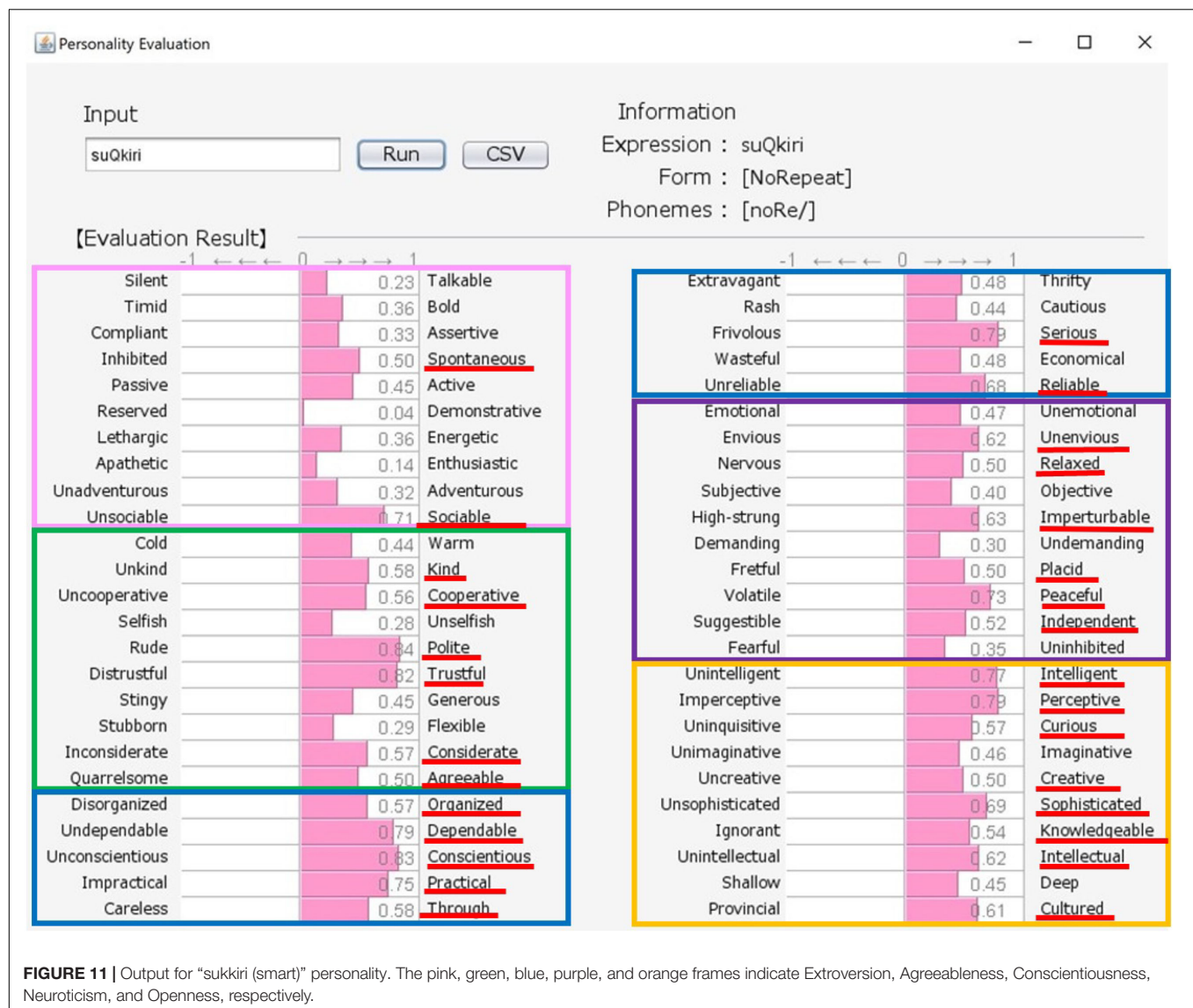


FIGURE 11 | Output for “sukiri (smart)” personality. The pink, green, blue, purple, and orange frames indicate Extroversion, Agreeableness, Conscientiousness, Neuroticism, and Openness, respectively.

secret and frequently shares information inappropriately with others (Badenoch et al., 2019). The SSW “gusu-gusu” describes a slow and silent, inactive personality (Osada et al., 2019). In addition, sound-symbolic features related to personality traits have been found in Indo-European languages, such as English and Spanish (Lindauer, 1990; Milán et al., 2013; Kawahara et al., 2015). It would be interesting to investigate the differences in personality categories among different languages, as our approach might be applicable to other languages that have SSWs. Future research could explore the reasons why some SSWs represent personality, and why these are particularly highly developed in Japanese.

CONCLUSION

In this paper, we focused on SSWs that can express complex aspects of personality traits, and constructed a system that

can convert a SSW into values in terms of 50 personality-related adjective pairs. This system can obtain information equivalent to the adjective scales using only a single word instead of asking many questions. The prediction accuracy of the system was tested by calculating the multiple correlation coefficients, and it was found that those of 48 personality-related adjective pairs exceeded 0.75. An evaluation experiment, in which participants rated the appropriateness of the system output, was also performed, and the result demonstrated the effectiveness of the system. We believe that this system can contribute to the field of personality computing.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

This study was reviewed and approved by the ethics committee of The University of Electro-Communications, Tokyo, Japan. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MS and JW conceived the experiments. MS performed the experiment. JW and KY carried out the data analyses. All authors discussed and interpreted the results, and contributed to drafts of this manuscript.

REFERENCES

- Allport, G. W., and Odbert, H. S. (1936). Trait-names: a psycho-lexical study. *Psychol. Monogr.* 47, i–171. doi: 10.1037/h0093360
- Badenoch, N., Purti, M., and Choksi, N. (2019). Expressives as moral propositions in Mundari. *Indian Ling.* 80, 1–17.
- Bloomfield, L. (1933). *Language*. New York, NY: Henry Holt.
- Bolinger, D. (1950). Rime, assonance and morpheme analysis. *Word* 6, 117–136. doi: 10.1080/00437956.1950.11659374
- Cattell, R. B. (1943). The description of personality: basic traits resolved into clusters. *J. Abnorm. Soc. Psychol.* 38, 476–506. doi: 10.1037/h0054116
- Costa, P. T., and MacCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional Manual*. Ukraine: Psychological Assessment Resources, Incorporated.
- Crystal, D. (1995). *The Cambridge Encyclopedia of the English Language*. New York, NY: Cambridge University Press.
- Das, K. G., and Das, D. (2016). “Developing lexicon and classifier for personality identification in texts,” in *Proceedings of the 14th Intl. Conference on Natural Language Processing*, (West Bengal: © 2016 NLP Association of India (NLP AI)), 362–372.
- Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. *Lang. Linguist. Compass* 6, 654–672. doi: 10.1002/Inc3.361
- Doizaki, R., Watanabe, J., and Sakamoto, M. (2017). Automatic estimation of multidimensional ratings from a single sound-symbolic word and word-based visualization of tactile perceptual space. *IEEE Trans. Hapt.* 10, 173–182. doi: 10.1109/toh.2016.2615923
- Fiske, D. W. (1949). Consistency of the factorial structures of personality rating from different sources. *J. Abnorm. Soc. Psychol.* 44, 329–344. doi: 10.1037/h0057198
- Goldberg, L. R. (1982). “From ace to zombie: some explorations in the language of personality,” in *Advances in Personality Assessment*, Vol. 1, eds C. D. Spielberger and J. N. Butcher (Hillsdale, NJ: Lawrence Erlbaum Associates), 203–234.
- Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychol. Assess.* 4:26. doi: 10.1037/1040-3590.4.1.26
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *Am. Psychol.* 48:26. doi: 10.1037/0003-066x.48.1.26
- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., et al. (2006). The international personality item pool and the future of public-domain personality measures. *J. Res. Personal.* 40, 84–96. doi: 10.1016/j.jrp.2005.08.007
- Hamano, S. (1998). *The Sound-Symbolic System of Japanese*. Tokyo: CSLI publications & Kuroshio.
- Hirsh, J. B., and Peterson, J. B. (2009). Personality and language use in self Narratives. *J. Res. Personal.* 43, 524–527. doi: 10.1016/j.jrp.2009.01.006
- Ito, J., Kanoh, M., Nakamura, T., and Komatsu, T. (2013). Editing robot motion using phonemic feature of onomatopoeias. *J. Adv. Comput. Intell.* 17, 227–236. doi: 10.20965/jaciii.2013.p0227
- Jespersen, O. (1922). The symbolic value of the vowel i. *Philologica* 1, 1–19.
- Kawahara, S., Shinohara, K., and Grady, J. (2015). “Iconic inferences about personality: from sounds and shapes,” in *Iconicity: East Meets West*, Vol. 3, eds M. K. Hiraga, W. J. Herlofsky, K. Shinohara, and K. Akita (Amsterdam: John Benjamins), 57–69. doi: 10.1075/ill.14.03kaw
- Köhler, W. (1929). *Gestalt Psychology*. New York, NY: Liveright.
- Komatsu, K., Sakai, K., Nishioka, M., and Mukoyama, Y. (2012). The gitaigo personality scale for description of self and others. *Japanese. J. Psychol.* 83, 82–90. doi: 10.4992/jjpsy.83.82
- Lindauer, S. M. (1990). The meanings of the physiognomic stimuli taketa and maluma. *Bull. Psychon. Soc.* 28, 47–50. doi: 10.3758/bf03337645
- Mairesse, F., Walker, A. M., Mehl, R. M., and Moore, K. R. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res.* 30, 457–500. doi: 10.1613/jair.2349
- McCrae, R. R., and Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *J. Personal. Soc. Psychol.* 52, 81–90. doi: 10.1037/0022-3514.52.1.81
- Milán, E. G., Iborra, O., de Córdoba, M. J., Juárez-Ramos, V., Artacho, M. A. R., and Rubio, J. L. (2013). The kiki-bouba effect: a case of personification and ideasthesia. *J. Conscious. Stud.* 20, 84–102.
- Nishioka, M., Komatsu, K., Mukoyama, Y., and Sakai, K. (2006). Japanese mimetic words “Gitaigo” used to describe personality: classification and semantics. *Japanese. J. Psychol.* 77, 325–332. doi: 10.4992/jjpsy.77.325
- Norman, W. T. (1967). *2,800 Personality Trait Descriptors: Normative Operating Characteristics for a University Population*. Abishekapatti: Department of Psychology, University of Michigan.
- Osada, T., Purti, M., Badenoch, N., Onishi, M., and Datta, D. P. (2019). *A Dictionary of Mundari Expressives*. Tokyo: ILCAA, Tokyo University of Foreign Studies.
- Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic Inquiry and Word Count: LIWC 2001*. Mahway, NJ: Lawrence Erlbaum Associates.
- Pennebaker, J. W., and King, L. A. (1999). Linguistic styles: language use as an individual difference. *J. Pers. Soc. Psychol.* 77, 1296–1312. doi: 10.1037/0022-3514.77.6.1296
- Ramachandran, V. S., and Hubbard, E. M. (2001). Synaesthesia—a window into perception, thought, and language. *J. Conscious. Stud.* 8, 3–34.
- Sakamoto, M., and Watanabe, J. (2016). Cross-modal associations between sounds and drink tastes/textures: a study with spontaneous production of sound-symbolic words. *Chem. Senses* 4, 197–203. doi: 10.1093/chemse/bjv078
- Sakamoto, M., and Watanabe, J. (2018). Bouba/kiki in touch: associations between tactile perceptual qualities and Japanese phonemes. *Front. Psychol.* 9:295. doi: 10.3389/fpsyg.2018.00295

FUNDING

This work was supported by JSPS KAKENHI Grant Number JP 15K12127.

ACKNOWLEDGMENTS

We thank Michelle Pascoe, Ph.D., and Sydney Koke, M.F.A., from Edanz Group (<https://en-author-services.edanz.com/ac>) for editing a draft of this manuscript. Part of this research was conducted following a master's thesis by Yusuke Kusaba in 2015.

- Sapir, E. (1929). A study of phonetic symbolism. *J. Exp. Psychol.* 12, 225–239. doi: 10.1037/h0070931
- Shinohara, K., and Kawahara, S. (2013). The sound symbolic nature of Japanese maid names. *Proc. JCLA* 13, 183–193.
- Tupes, E. C., and Christal, R. E. (1961). *Recurrent Personality Factors Based on Trait Ratings. Technical Report, USAF*. San Antonio, TX: Lackland Air Force Base.
- Uchida, T. (2005). Effects of intonation contours in speech upon the image of speakers' personality. *Japanese J. Psychol.* 76, 382–390. doi: 10.4992/jjpsy.76.382
- Vinciarelli, A., and Mohammad, G. (2014). A survey of personality computing. *IEEE Trans. Affect. Comput.* 5, 273–291. doi: 10.1109/taffc.2014.2330816
- Yarkoni, T. (2010). Personality in 100,000 words: a large-scale analysis of personality and word use among bloggers. *J. Res. Personal.* 44, 363–373. doi: 10.1016/j.jrp.2010.04.001

Conflict of Interest: JW is employed by the NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation as a research scientist conducting basic scientific research on human sensory processing. There are no patents, products in development, or marketed products to declare. This does not alter the authors' adherence to the journal's policies on sharing data and materials.

The remaining 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.

Copyright © 2021 Sakamoto, Watanabe and Yamagata. 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.



Developing an Error Map for Cognitive Navigation System

Atsushi Ito^{1,2*}, Yosuke Nakamura^{2†}, Yuko Hiramatsu¹, Tomoya Kitani³ and Hiroyuki Hatano⁴

¹ Faculty of Economics, Chuo University, Tokyo, Japan, ² Faculty of Engineering, Utsunomiya University, Utsunomiya, Japan,

³ Faculty of Engineering, Shizuoka University, Shizuoka, Japan, ⁴ Faculty of Engineering, Mie University, Tsu, Japan

OPEN ACCESS

Edited by:

Péter Baranyi,
Széchenyi István University, Hungary

Reviewed by:

Jozsef Katona,
University of Dunaújváros, Hungary
György Molnár,
Budapest University of Technology
and Economics, Hungary

*Correspondence:

Atsushi Ito
atc.00s@g.chuo-u.ac.jp

†Present address:

Yosuke Nakamura,
OKI Software Co., Ltd., Saitama,
Japan

Specialty section:

This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

Received: 31 January 2021

Accepted: 21 March 2021

Published: 04 May 2021

Citation:

Ito A, Nakamura Y, Hiramatsu Y,
Kitani T and Hatano H (2021)
Developing an Error Map for Cognitive
Navigation System.
Front. Comput. Sci. 3:661904.
doi: 10.3389/fcomp.2021.661904

Global Navigation Satellite System (GNSS) positioning is a widely used and a key intelligent transportation system (ITS) technology. An automotive navigation system is necessary when driving to an unfamiliar location. One difficulty regarding GNSS positioning occurs when an error is caused by various factors, which reduces the positioning accuracy and impacts the performance of applications such as navigation systems. However, there is no way for users to be aware of the magnitude of the error. In this paper, we propose a *cognitive navigation system* that uses an *error map* to provide users with information about the magnitude of errors to better understand the positioning accuracy. This technology can allow us to develop a new navigation system that offers a more user-friendly interface. We propose that the method will develop an error map by using two low-cost GNSS receivers to provide information about the magnitude of errors. We also recommend some applications that will work with the error map.

Keywords: Global Navigation Satellite System, error map, cooperative positioning, intelligent transport system, CogInfoCom

1. INTRODUCTION

Many types of navigation services, such as applications on smartphones and automotive navigation systems, are becoming popular. These services play an important role in intelligent transportation systems (ITSs). However, they require highly accurate positioning, especially in urban areas. One technology that is important for providing location information such as latitude and longitude is the Global Navigation Satellite System (GNSS). Position information determined by GNSS usually is inaccurate by a few meters or more, and research has been performed to reduce the error. One promising technology for acquiring accurate location data is real-time kinematic (RTK) positioning (Sakai, 2003). In the best-case scenario, RTK positioning can provide location information with an error of only a few centimeters. Owing to its low cost and small size, the RTK positioning device is expected to be used widely. However, its positioning accuracy is not good where signal reception from satellites is unstable, especially in urban canyons. RTK accuracy depends on the environment, such as when buildings shadow signals from a GNSS satellite, so it is difficult to realize accurate navigation with RTK positioning in a city's downtown area. One solution for solving the challenges in an urban environment is to select a satellite to eliminate signal shading by buildings. An elevation angle mask (Misra and Enge, 2001) is a technique that provides accurate position information in cities by employing satellites that exist at higher-elevation angle spaces. However, such methods for selecting satellites sometimes reduce the number of usable satellites, so accuracy does not increase as expected. Therefore, the accuracy of GNSS positioning in urban areas is challenging in general.

However, we believe that there is an alternative approach wherein applying a cognitive methodology provides user satisfaction for services such as car navigation. The main stream of the research on GNSS is increasing accuracy. However, we sometimes have a problem of car navigation because of low accuracy of position. If the positioning accuracy is not good because of the conditions, we expect that a navigation application provides an optimal solution. For example, if the accuracy of position is not good, a navigation system can lead a user to a location where the accuracy is better. We would like to call such navigation system as a *cognitive navigation* system. Therefore, we would like to apply the *Cognitive Infocommunications (CogInfoCom)* approach (Baranyi and Csapo, 2012; Baranyi et al., 2015). This idea extends human cognitive capabilities and would even enable life support.

Figure 1 shows the CogInfoCom concept for a location-based service (LBS). The original CogInfoCom architecture is shown in **Figure 1** (left). A communication device offers an opportunity for users to interact with the environment by using a network and an autonomous cognitive system with sensors and Internet of Things devices. **Figure 1** (right) depicts one implementation of the CogInfoCom system to an LBS, namely, an automotive navigation system. Clearly, LBSs are one of the most important services and are strongly affected by situations. In this implementation, we introduce an *error map* (**Figure 2**) that offers a level of error for GNSS positioning to improve navigation if the GNSS signal is weak or noisy. CogInfoCom makes an environment intelligent and provides information automatically.

In the next section, we propose a new technique that increases the accuracy of GNSS positioning and the usability of LBSs by providing the error level of a specific location. We mention the difficulty of GNSS positioning and related works such as techniques for multipath shadowing environments or multi-GNSS receivers. Section 3 discusses the objectives of this research, and section 4 presents the pre-examination results to explain the characteristics of the GNSS positioning error. In section 5, details of the proposed method are discussed. The proposed methods are

evaluated in section 6, and examples of how to apply the error map are proposed in section 7. Finally, section 8 concludes and discusses further studies.

2. PROBLEMS AND RELATED WORKS

2.1. Characteristics of the GNSS Error

As described in the previous section, one of the issues related to reducing the accuracy of GNSS is errors. A GNSS positioning error can occur for several reasons. Such an error can be caused by any of the following: the satellite's clock, the orbit of a satellite, noise from the convection of air in the ionosphere, noise of a GNSS receiver, or multipaths (Sakai, 2003). An error that is caused by a clock's satellite or the satellite's orbit relates to the satellite itself, so it is possible to reduce the error if more than two receivers obtain a signal from the same satellite. In widely used techniques such as the Saastamoinen model (Saastamoinen,

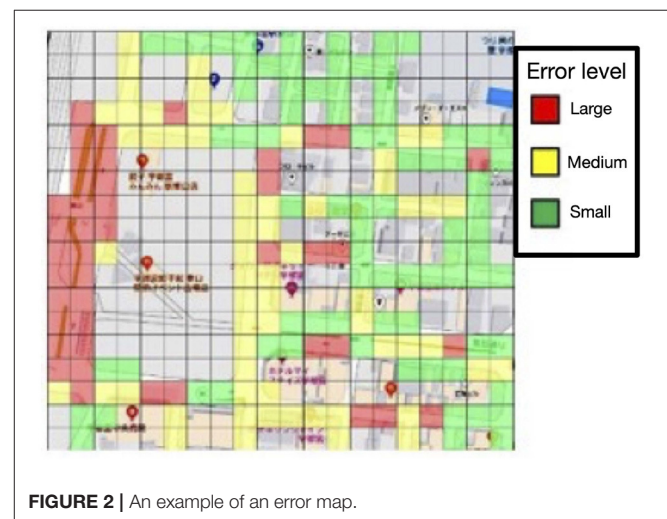
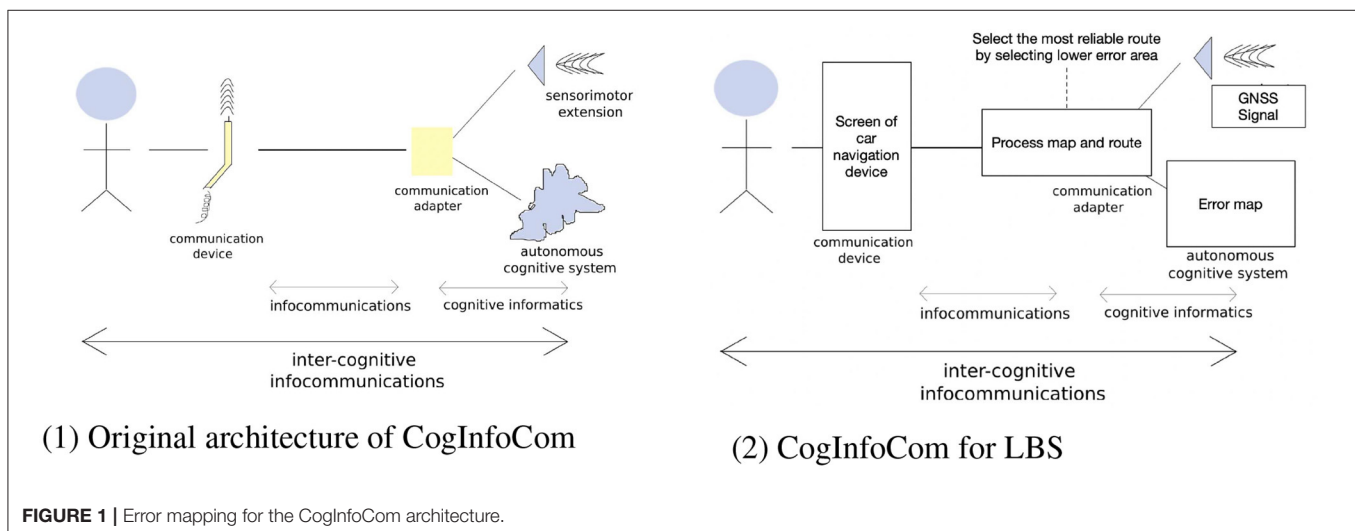


FIGURE 2 | An example of an error map.



1973a,b,c) and the Klobuchar model (Klobuchar, 1987), an error caused by noise from the GNSS receiver depends on the baseline (Satirapod and Chalermwattanachai, 2005). Therefore, it is possible to remove the miscalculation as a common error if there are two neighbor receivers. The remaining error depends on shadows, which cause reflections and increase the signal paths from GNSS satellites, and such reflections interfere with the original signal. Such an error varies, depending on the environment, and is difficult to predict. Also, the objects' shadows reduce the number of GNSS satellites from a particular receiver that secures the signal and increases error. The shadowing environment causes errors that are difficult to remove, and the levels of error depend on the environment. Therefore, it is important to handle these types of errors for an LBS.

2.2. Coordinated Positioning

One approach for reducing errors is coordinated positioning, which minimizes miscalculations that are caused by sharing information among several receivers. The idea is similar to RTK positioning and removing the common receiver errors. The coordinated positioning can detect multipaths and eliminate the satellite that causes them (Osechas et al., 2015) to increase accuracy. A Differential Global Positioning System (DGPS) is one type of coordinated positioning. The reference station distributes correction information to neighboring receivers. Accordingly, a simplified DGPS is proposed (Miyata et al., 1996; Miyata and Sakitani, 1997). This system offers accurate positioning data by processing the cross-correlation of positioning data of two receivers. There is also a method for grouping the characteristics of GNSS receivers, such as an error-reduction method for the receivers that move together (Odaka et al., 2011) and to reduce multipaths by using numerous antennas on an automobile (Kubo et al., 2017).

2.3. Error Correction by a Map

A map is an important parameter for increasing the positioning accuracy and provides both correction and height information (Iwase et al., 2013). The information provided by a map is increasing, such as 3D maps. We also propose a method to provide DGPS correction data by combining map information and the location of an automobile (Rohani et al., 2016).

3. ESTIMATING THE ERROR OF GNSS

In this section, we present the results of an experiment that considers a method to determine an estimated position error. As explained in the previous section, it is difficult to eliminate errors caused by a multipath. Table 1 presents the methods used in previous studies. These methods used the pseudorange. However, for a low-cost GNSS device such as a smartphone, the positioning function is a black box, and it is difficult to see the pseudorange information. Therefore, it is necessary to use only position data, with no internal data from the GNSS receiver. A way to use the group characteristics of GNSS receivers is to apply only the position data. However, because it is assumed to be stationary, it requires developing an error-reduction method that uses only the mobile receiver's position data. We have been developing

TABLE 1 | Related works.

Method	Required data
Multi-path detection using GNSS application	Pseudorange, carrier phase
Using group characteristic of GPS receiver	Position data without movement
Multi-path elimination by using a map	Pseudorange, height data
Collection of information by multiple receivers and map-matching	Pseudorange, map

some methods to reduce errors that relate to multipath under the restrictions mentioned. In the remainder of this section, we will discuss our methods.

3.1. Method 1: Using the Common Temporal Error

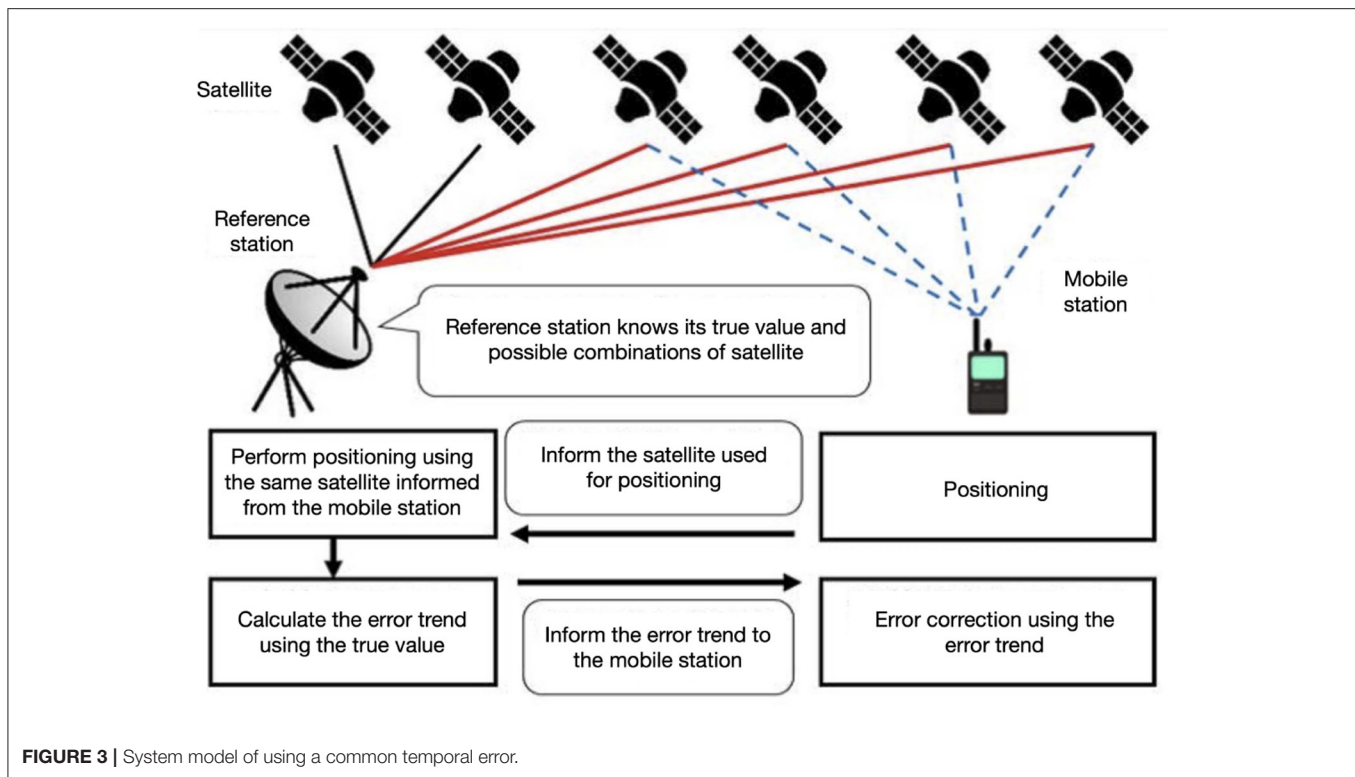
We studied a method that focuses on how a common temporal error occurs in nearby receivers for the GNSS (Nakamura, 2018). The delay due to the convection of air in the ionosphere causes the main GNSS positioning error. If the distance between two receivers is small, both have the same effect caused by the delay due to the convection of air in the ionosphere and may have a common error. However, if each receiver obtains a signal from a different GNSS satellite, the pseudorange should vary and have different error patterns. Therefore, it is necessary to use the same GNSS satellite. To prove the effectiveness of this idea, we designed a system model (Figure 3) and performed a trial. As a result, accuracy improved by 0.5–1.6 m.

3.2. Method 2: Using a Bias of the Positioning Error by Multipath

This idea uses an error bias caused by multipath between neighboring receivers. In our previous research (Kitani et al., 2012), the error of neighboring receivers is biased. In this method, two receivers were placed on top of the roof of a car with the same distance, and the error of position was observed. This automobile was moving straight, and the environment alternates between not being in shadow and being in shadow. If the observed error of the two receivers fluctuates widely, both are affected for the same reason. In this case, we can understand that both receivers are affected by multipath, and we can add a corrected point by using a previously accurate result in the shadowed environment.

3.3. Method 3: Error Map

Methods 1 and 2 were designed to reduce errors. However, our idea is different in that if we know the location is shadowed and what the possible error level is, we can adapt the situation. We would like to propose a new idea, an error map, for that purpose. Figure 2 shows an example of an error map in which different colors display the multipath's level of error. For example, if we know that the location has a poor error level, we can either change a shadow mask to eliminate the low-accuracy satellite or select a route with a better error level.



3.4. Comparing the Methods

In what follows, we compare the three proposed methods.

- Method 1 uses a common temporal error and has value for correcting an error by using position information. However, this technique applies to locations without shadowing. Also, if both the reference station and the mobile terminal are in a shadowy condition, it is possible to remove the error. The reference station is usually located at a position without a shadow, so it is not useful in the usual case.
- Method 2, which uses the bias of the positioning error by multipath, is useful in a limited situation, since a case with a similar multipath trend is rare.
- In Method 3, the error map is not useful for removing the error. This method can change the shadow mask to reduce errors. However, when using this technique, a participant can choose various ways that are not affected by the GNSS positioning error. The error map can also provide information to improve the infrastructure, such as a beacon to provide location information in an urban area, such as a dynamic traffic map (Watanabe et al., 2020). We believe that this method provides many possibilities to develop LBSs.

Accordingly, we selected the Method 3 error map because of the research target since the method can be applied to the LBSs. The following are the objectives of this research.

- Requirement 1: To develop a function to estimate the error, with a 1 m target for 90% of the cases

- Requirement 2: To cultivate a function to decide the error level, with a target of 90% success rate at the error-level decision.

4. PRE-EXPERIMENT

We performed the following two pre-experiments to evaluate the error size by using two GNSS receivers and their position.

- Pre-experiment 1: observing the trend of positioning results affected by multipath measured in both shaded and unshadowed environments by using two GNSS receivers (stationary)
- Pre-experiment 2: the same test but in the moving case.

4.1. Evaluating Pre-experiment 1

We performed pre-experiment 1 to observe how the trend of positioning results of moving two receivers that are affected by multipath was measured under shaded and unshadowed environments. The experimental factors are shown in **Table 2** (1). The reference station was not shadowed. The user's receiver was shadowed and affected by multipath and was located close to the large building, with a distance of approximately 20 m.

In this experiment, true value is defined as the average of the fixed solution. The distance between the two receivers at the unshadowed environment was 0.947 m, whereas that in the shadowed environment was 0.247 m. Since the real distance was 1 m, the result of the shadowed environment was not good. Also, in the unshadowed environment, some epochs do not have an output. Although there were some

TABLE 2 | Experimental factors for pre-experiment 1 and 2.

Factor	Value
(1) Experimental factors for pre-experiment 1	
Date	June 2, 2018
Time	7:40 a.m.–10:40 a.m.
Location	Reference (unshadowed): the roof of the six-floor building, User receiver (shadowed): near the tall building
Device	Reference: EVK-M8T, User Receiver: u-blox NEO-M8T
Antenna	Tallysman TW2710
Frequency	1 Hz
Distance between devices	1 m
GNSS satellite	GPS, BeiDou
(2) Experimental factors for pre-experiment 2	
Date	July 8, 2018
Time	21:40–21:52
Location	Reference (unshadowed): the roof of the six-floor building in the Utsunomiya University campus; User receiver (moving): in a Utsunomiya University campus
Device	Reference : EVK-M8T, User : u-blox NEO-M8T
Antenna	Tallysman TW2710
Frequency	1 Hz
Distance between devices	1 m
GNSS satellite	GPS, BeiDou, Quasi-Zenith Satellite System (QZSS)

errors, they were not as severe. **Figure 4** (1, 2) shows the error in the shadowed (light color) and unshadowed (dark color) environments at point a, whereas **Figure 4** (3, 4) shows the error at point b. If there is no miscalculation, the difference should be 0. The error in the shadowed environment is larger than that in the unshadowed one. The error trend is changing slowly and extensively. From this experiment, we predict that in the shadowed environment, the noise increases. So it may be possible. The shadowed environment has an error by multipath. Also, since the miscalculation trend is changing slowly and significantly in the shadowed environment, we believe that the effect of satellite constellation is more significant than that in the unshadowed environment.

4.2. Evaluating Pre-experiment 2

We performed the next experiment to observe the trend of the positioning results of moving two receivers affected by multipath. We examined the relationship between the average error of the position of each receiver and the distance from the receivers. We set two receivers on a push car, which moved to Utsunomiya University campuses and measured the positions. The distance between receivers was 40 cm. The moving route was selected for motion in the shadowed area. **Table 2** (2) shows the experimental factors. We used a fixed RTK solution as the true value and calculated the positioning error as the distance using Hubeny's formula (Vincenty, 2013) and the longitude and latitude of the receiver's true value and position.

We define the errors in two receivers as e_1 and e_2 , and the average of the errors is $\bar{e} = \frac{e_1 + e_2}{2}$ [m]. Then, we defined the distance between two receivers as \tilde{r} [m], as we would like to evaluate the relationship between the average positioning error (\bar{e}) and the distance between two receivers (\tilde{r}). This examination was performed at the epoch, where the positioning error (e_1 , e_2) and distance between receivers (\tilde{r}) were acquired.

Figure 5 (left) shows how if the distance between receivers (\tilde{r}) were increased, the average positioning error of two receivers (\bar{e}) was increased. **Figure 5** (right) is an enlargement of **Figure 5** (left), where the distance between receivers (\tilde{r}) is between 0 and 10 m. This figure shows that there are many points scattered in the positive direction of the vertical axis, but not in the negative direction. **Figure 5** (left) and (right) show the distance between receivers (\tilde{r}) and the average positioning error of two receivers (\bar{e}), which has a positive correlation because if \bar{e} becomes larger, \tilde{r} should be larger.

The average positioning error of two receivers (\bar{e}) requires the true value, although the distance between receivers (\tilde{r}) can be calculated from the positioning data of the receiver. Therefore, we believe that it is possible to calculate the average of the positioning error of two receivers (\bar{e}) from a distance between receivers (\tilde{r}) by using two low-cost GNSS receivers. If we can obtain the average of the positioning error of two receivers (\bar{e}), we can use that data to develop an error map.

In the next section, we explain the method to calculate the estimated positioning error.

5. THE PROPOSED METHOD FOR CALCULATING THE ESTIMATED POSITIONING ERROR FOR THE ERROR MAP

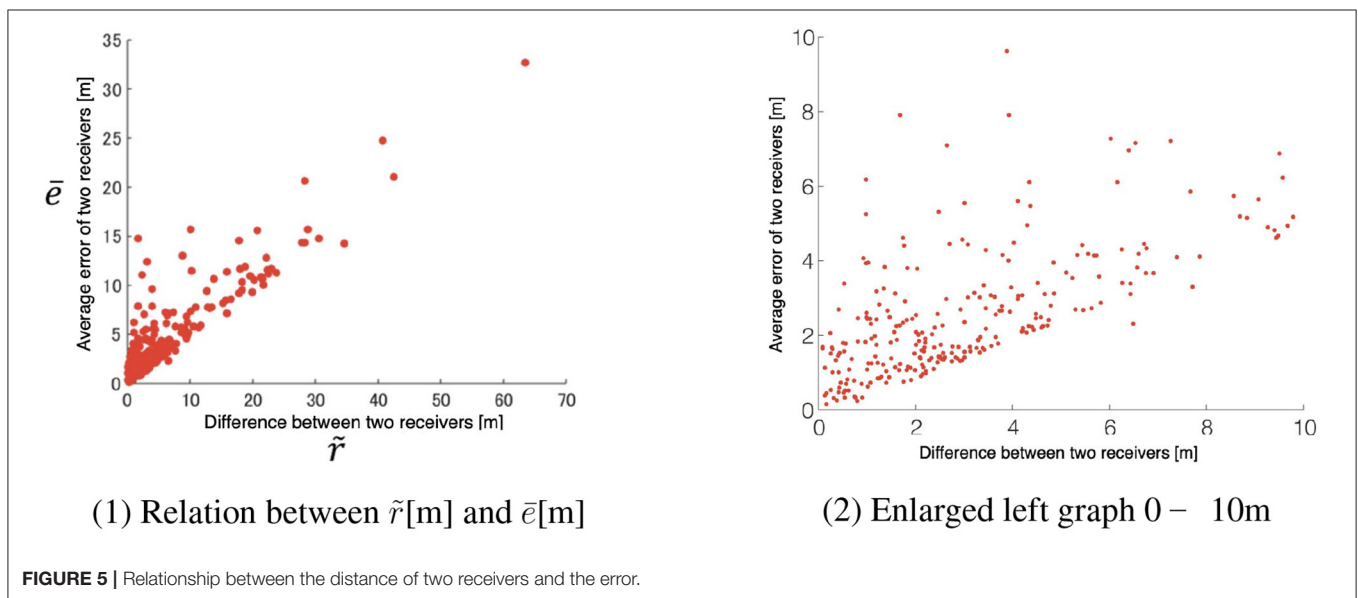
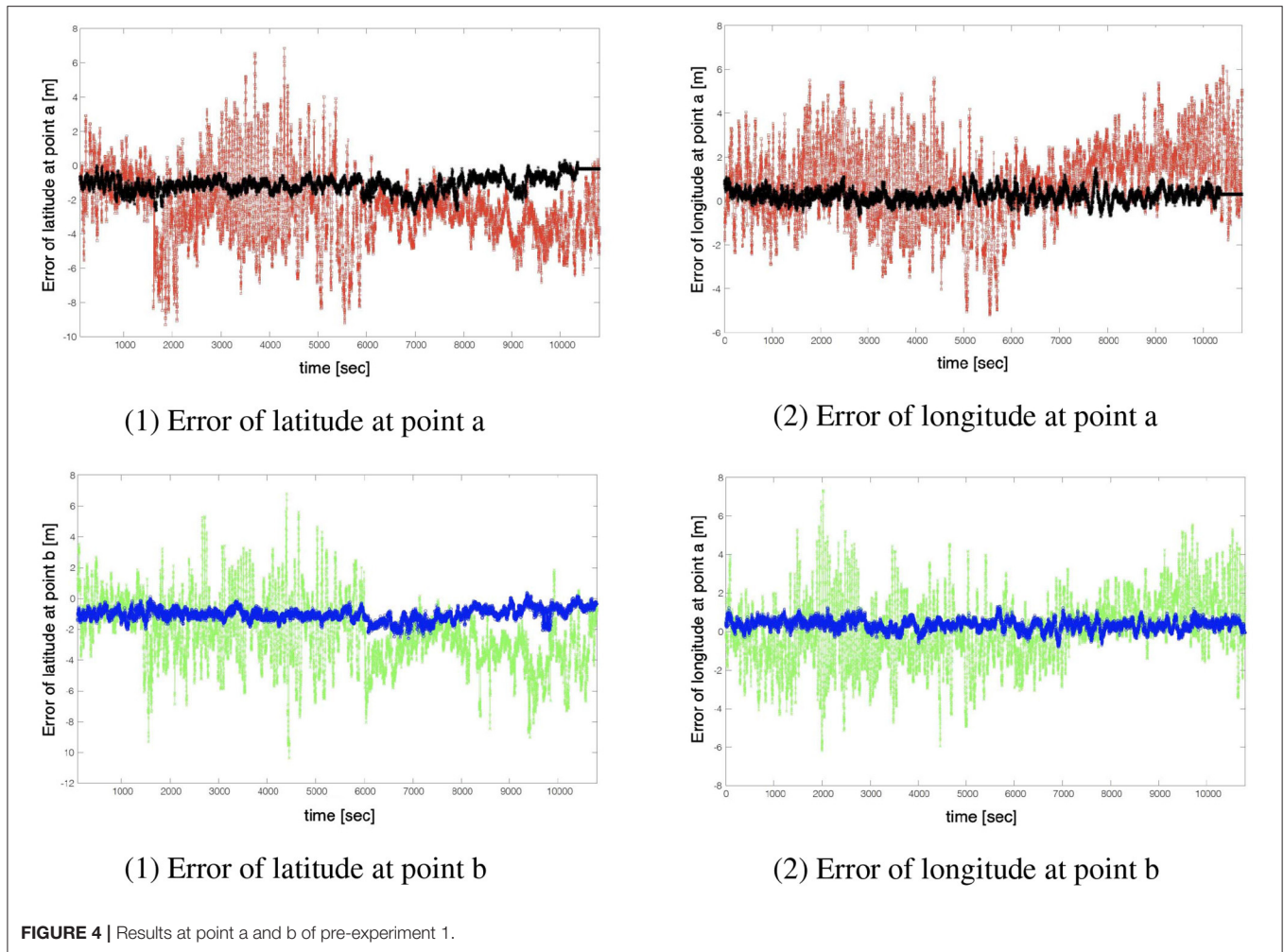
This section presents the details of our proposed method for calculating the estimated positioning error for the error map.

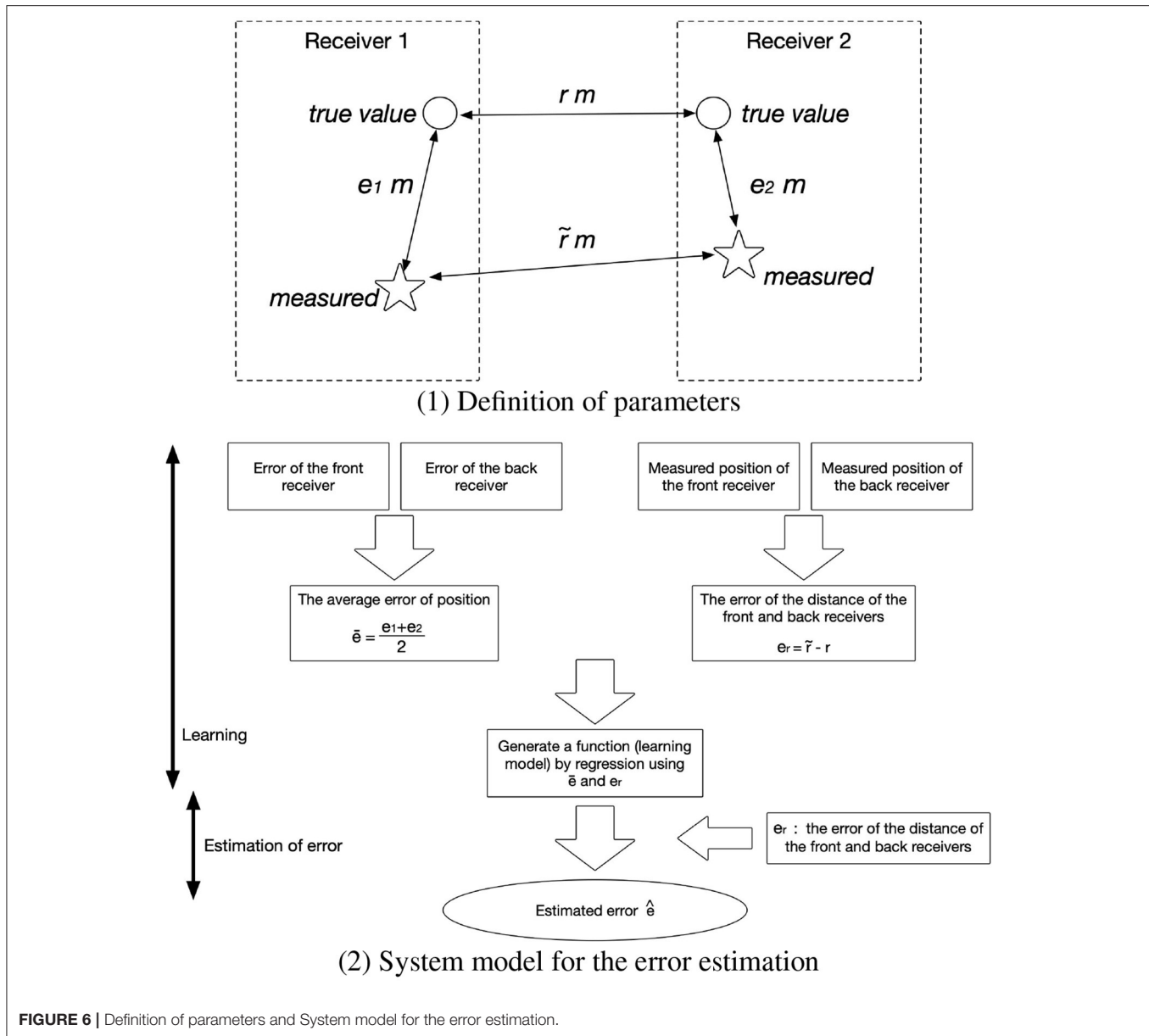
5.1. Error Estimation

Figure 6 (1) depicts the parameters and **Figure 6** (2) illustrates the system model for the error estimation.

Let us assume there are two receivers located in front of and behind the roof of a car. We use the following parameters:

- e_1 [m]: positioning error between the true value and the standalone positioning at the front receiver
- e_2 [m]: positioning error between the true value and the standalone positioning at the back receiver
- \tilde{r} [m]: the measured distance between two receivers acquired from standalone positioning
- r [m]: the distance between two receivers acquired from the true value
- \bar{e} [m]: positioning error of two receivers [calculated by Formula (1)]
- e_r [m]: error of the distance of two receivers [calculated by Formula (2)].





$$\begin{aligned} \bar{e} &= \frac{e_1 + e_2}{2} [m] \\ e_r &= \tilde{r} - r [m] \end{aligned} \quad \begin{matrix} (1) \\ (2) \end{matrix}$$

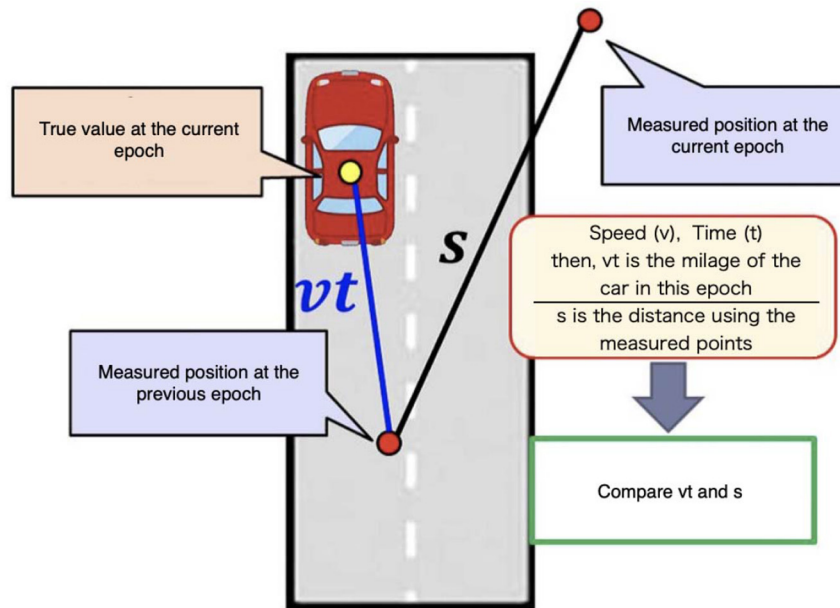
We obtained the error estimation function by performing a regression analysis, with the front-to-back receiver distance error e_r since the x-axis and the front-and-back positioning error mean \bar{e} as the y-axis. From the shape of **Figure 5** (left), we performed linear regression and trained the intercept to become 0. Subsequently, we obtained the relation between e_r and \bar{e} . Using this function, we can obtain e_r from \tilde{r} using Formula (2). We can then estimate \bar{e} .

5.2. Deciding on the Magnitude of Error

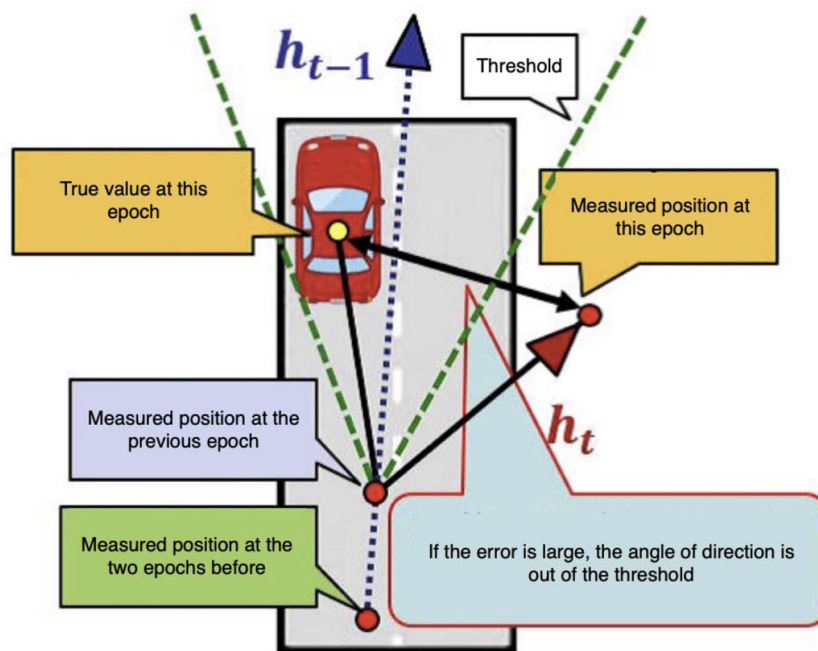
We discuss the error level evaluation method (i.e., large or small) in this subsection. We also propose methods for deciding on the error level based on the discussion in the previous section. We decide that “the error is large” if the error is larger than \tilde{r} ; otherwise, “the error is small.” However, this simple method sometimes causes an unexpected problem; hence, we would like to discuss this in detail.

5.2.1. Method Using Mileage and Distance Between Positioning Points

The method explained in the previous section has some problems when deciding on the error level. For example, the positioning point receives an error in the same direction and



(1) Using mileage and distance between the positioning points

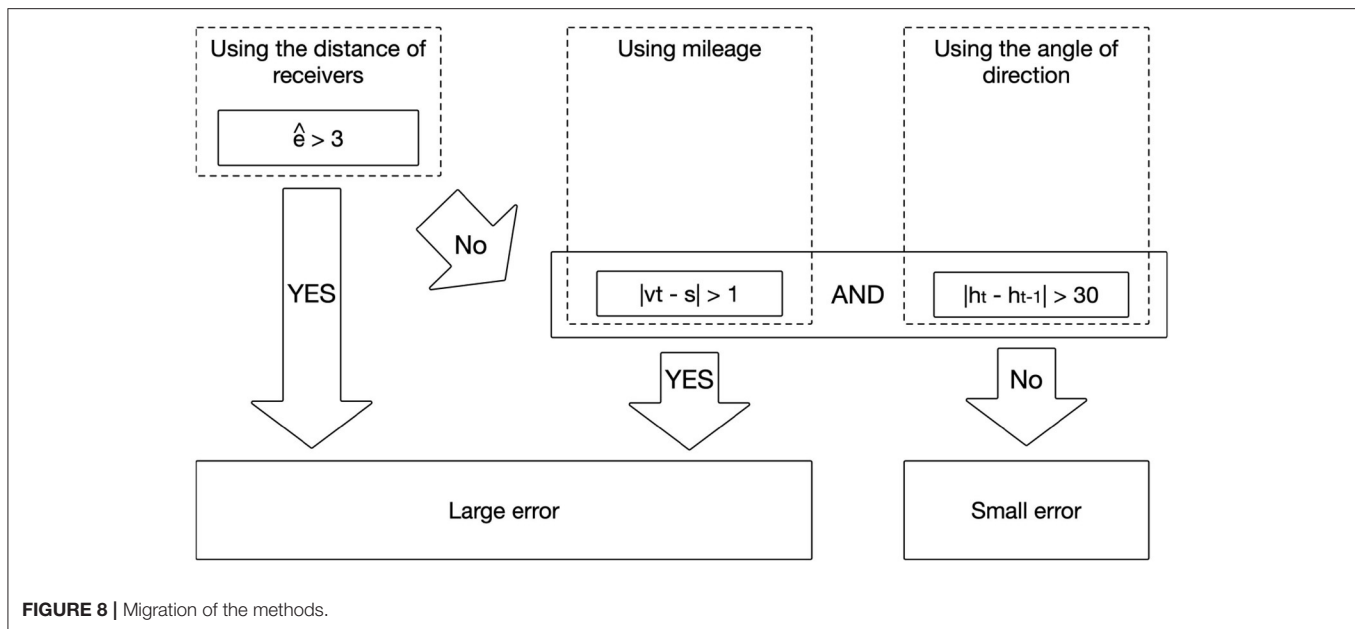


(2) Using the angle of direction

FIGURE 7 | Error level estimation methods.

becomes close to the receiver distance's true value, despite receiving the error. Therefore, the distance calculated from the current time t , the time t' of the previous epoch, and the speed v are the pseudo-true values. We can then estimate the magnitude of error by taking the difference between the

current positioning point and the distance s obtained from the previous epoch's positioning point [Figure 7 (1)]. This value is the difference between the distance that should have been traveled and the distance that was traveled. Thus, we can decide on this considering the following: the error is large if the



value is large, and the error is small if the value is small. Therefore, the problem mentioned at the beginning of this section can be solved in many cases. However, this solution still has issues if the positioning points appear in the vehicle's opposite direction.

5.2.2. Method Using the Angle of Direction

We introduce a method for reinforcing the proposed method by using the angle of direction to solve the previous section's problem. This method is used as an estimation reinforcement when the vehicle is going straight on a straight road. First, the vehicle's traveling direction angle is obtained using the past two stable positioning points. Next, the threshold of the angle of the vehicle's traveling direction is set. Finally, we determine if the positioning error is small by determining whether the current epoch's positioning point is within the threshold of that angle. **Figure 7 (2)** presents an example. As in the lower-left part, the positioning point of the previous epoch and the vehicle's traveling direction angle determine whether the current epoch's positioning point is within the threshold. In this case, it was not within the threshold; thus, it is a "large error."

5.2.3. Migration of the Methods

We migrate the proposed methods in this section (**Figure 8**). First, the distance between the two receivers is judged, and we judge whether \tilde{r} exceeds 3[m]. If $\tilde{r} > 3$, we judge this as a "large error." If $\tilde{r} < 3$, we judge this as a "small error." The methods using mileage and angle of the vehicle's traveling direction are applied. If both methods answer "large error," the result should be "large error"; otherwise, the result is "small error."

6. EVALUATION OF THE PROPOSED METHODS

This section explains the proposed methods' evaluation results using a car with two receivers on the roof. The calculation was performed after obtaining the data.

6.1. Evaluation of the Error Estimation Method

Table 3 (1) lists the experiment factors of evaluation of the error estimation method.

Two low-priced receivers (u-blox EVK-M8T) were installed on a car roof [**Figure 9 (1)**]. The distance of the receivers was 1 m. The true value of the position was detected using POS LV 620 from Trimble. POS LV 620 can perform positioning with an accuracy of several centimeters by integrating an inertial measurement unit (IMU) and a DMI (odometer) in addition to a highly accurate GNSS receiver antenna.

Figure 9 (2) depicts the route traveled to obtain the experimental data. After leaving Utsunomiya University, Yoto Campus, the car passed through Utsunomiya Station and went to Tobu Utsunomiya Station before returning to Yoto Campus.

The regression analysis data were calculated as the error of the distance of the two receivers (e_r) and the receiver average positioning error (\bar{e}) from the data measured at 1:48:39 p.m.–2:40:36 p.m.. The data used for the evaluation were measured at 11:14:20 a.m.–01: 00: 11: 56 p.m.. We used e_r of the distance between the front and rear receivers. The error estimation value \bar{e} output by the function discussed in the previous section was compared with the true value. The estimated positioning error was evaluated for accuracy. The estimated error (\bar{e}) obtained by the proposed method was subtracted from the positioning error (e_1) of the front receiver to obtain the absolute

value (i.e., $|\hat{e} - e_1|$) summarized as a frequency in **Figure 10** (left). **Figure 10** (right) depicts a similar result for the back receiver. The value of $x = 10$ [m] is the frequency, including all 10 m or more errors. The cumulative frequency is also displayed. A comparison of **Figure 10** (left) and **Figure 10** (right) shows that more than 80% of the estimated error was within 3 m (both within the front and back). Furthermore, in a common part, the epochs with errors of 1 and 2 m are the most in error estimation. The results of **Figure 10** (left) are shown in **Table 4** as values. The number of epochs is very small when the error is larger than 3 m. The results showed that 80% or more of the estimated error (\hat{e}) showed an error of 3 m or less from the actual positioning error. We think that the reason for the error of a few meters is the processed regression analysis. The intercept became 0. We also think that no perfect correlation exists in the data, as shown in **Figure 5** (left). If the error e_r in the distance between the front and rear receivers became closer to 0, it became inaccurate in the former case. We think that the reason for this is that the learning process was performed. The intercept became 0. Furthermore, the latter led to the method's performance degradation because the function's incompleteness caused it due to the lack of a perfect correlation. This method can also provide a quantitative error amount. Our target was the estimated error of 1 m; however, the result that satisfied this target was approximately 20%. Hence, at this moment, our method is not suitable for a service that requires a strict error level. Our method can be used for services that do not require strict accuracy, such as notifying the user of how reliable the current positioning is.

6.2. Evaluation of the Method Using Mileage and Distance Between Positioning Points

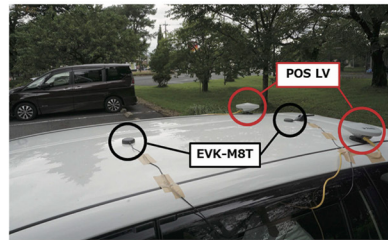
The evaluation was performed to confirm how much the error can be judged by the proposed method and how much the positioning accuracy can be improved using the "error map." We used the same devices that we used in the previous evaluation. **Table 3** (2) presents the experiment factors. In this experiment, we prepared two types of cases: one for the shadowed environment and one for the not-shadowed environment. The non-shadowed environment was set up near Utsunomiya University Yoto Campus. This area has a few high shields to the left and right of the road, and a stable positioning is possible. The shadowed environment was set up at the rotary at the west exit of Utsunomiya Station. There are many tall buildings around this area, and a pedestrian bridge covers the zenith direction. The evaluation was performed by using the method for determining the magnitude of the error and comparing it with the magnitude of the front receiver's actual error amount. After determining the magnitude of the error using this method, we confirmed whether the positioning accuracy could be improved by selecting satellites using the elevation mask. In the case of a "large error," we predicted the positioning accuracy could be improved by selecting a high-elevation satellite that does not cause a multi-path. In the case of a "small error," we predicted the positioning accuracy could be improved by using more

TABLE 3 | Experiment factors of the proposed method.

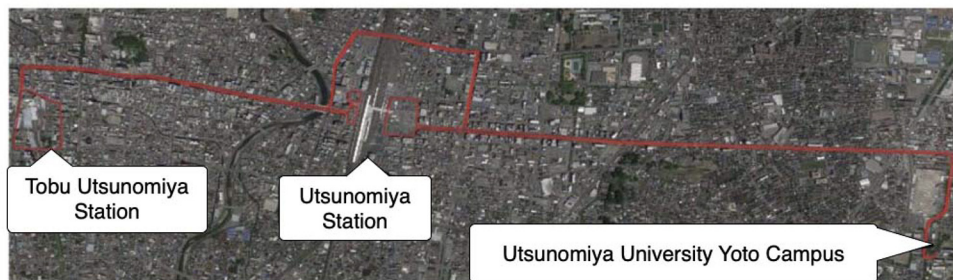
Factor	Value
(1) Experiment factors of evaluation of the error estimation method	
Date	Sep. 11, 2018
Time	AM11:14:20–PM 0:11:56, PM1:48:39–PM2:40:36
Location	Reference: the roof of the six-floor building in the campus of Utsunomiya University
Device	Reference : EVK-M8T, User : u-blox NEO-M8T
Antenna	Tallysman TW2710
Frequency	1 Hz
Distance between devices	1 m
GNSS satellite	GPS, BeiDou, ZSS
(2) Experiment factors of evaluation of the method using mileage and distance between positioning points	
Date	Sep. 11, 2018
Time	AM11:17:30–PM 11:20:47, AM11:35:04–PM11:38:06
Location	Not shadowed: Yoto Campus of Utsunomiya University; Shadowed: Utsunomiya Station West Gate
Device	Reference : EVK-M8T, User : u-blox NEO-M8T
Antenna	Tallysman TW2710
Frequency	2 Hz
Distance between devices	1 m
GNSS satellite	GPS, BeiDou, QZSS

satellites in such a way that the weak elevation mask would be useful. We want to verify the error map effectiveness. **Table 5** (1) describes the result of the non-shadowed environment. The success rate of the judgment was 97.6%. The result showed that the error size was successfully judged with high accuracy. Almost all actual errors showed a "small error" in the epoch, which is the reason why the judgment was performed well by the method. **Table 5** (2) describes the results of the shadowed environment. The judgment success rate was 72.1%, which was not highly accurate. We verified to what extent the elevation mask effect was obtained with this success rate. To calculate the positioning error, we set the elevation mask to 15° (general elevation mask value) when the judgment was "small error" and to 30° when the judgment was "large error." The results were then compared to those when the mask was set to 15° all the time.

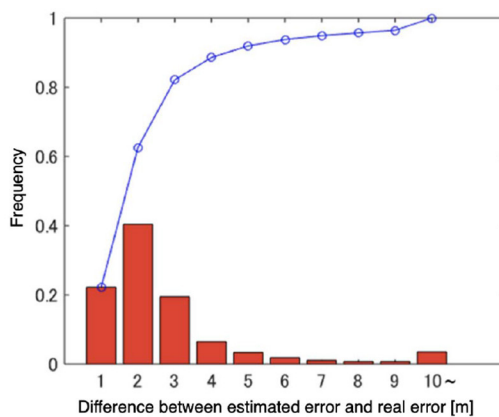
The average positioning error of fixed elevation mask (15°) was 5.78 m, and that of variable elevation mask was 5.50 m. This result shows that when the elevation mask was adjusted according to the judgment result, the average positioning error was improved by approximately 0.3 m compared to when the elevation mask was always set to the same value of 15°. By applying this method, the positioning accuracy was improved for 59.5% of the epoch. The loss of six of 273 epochs can be prevented compared to when the elevation mask was always set to 15°. In the non-shadowed environment, the error size was successfully discriminated by the method, and the discrimination was very



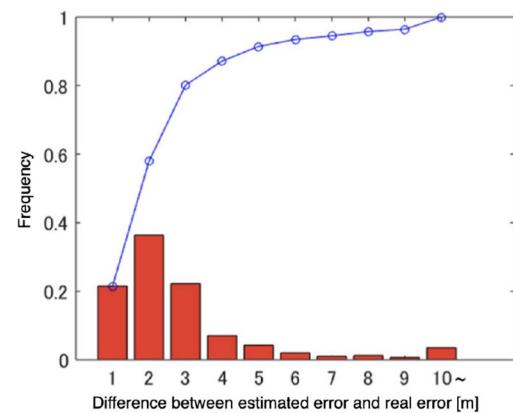
(1) Devices on the car roof



(2) Test route

FIGURE 9 | Devices on the car roof and test route.

(1) Frequency of error (front)



(2) Frequency of error (back)

FIGURE 10 | Frequency of error.

high at 97.6%. However, in the shadowed environment, the discrimination was at 72.1%. In this environment, the judgment's success rate decreased because the parts discriminated against as "small error" and as "large error" were mixed. However, as a result of setting the elevation mask using this judgment result, we confirmed that the positioning accuracy was improved, albeit with a small value of 0.3 m. The positioning accuracy did not improve much because only a few epochs can significantly improve 1 m or more when the 15° and 30° elevation masks were applied. Therefore, we believe that a great effect can be expected in an environment where the positioning accuracy can be greatly improved by adjusting the elevation mask. Moreover,

distinguishing between regions where positioning is stable and where it is unstable is possible. Hence, this method can be used to construct an error map.

The estimation error method could be used to estimate the positioning error. However, a precise error amount cannot be obtained; therefore, we need to select a service that can use this method. We also confirmed that the method using mileage and distance between positioning points could be used for the magnitude of error judgment, especially in the non-shadowed environment, and displays a high error judgment rate. On the contrary, the judgment accuracy in the shadowed environment was lower. We evaluated the effect of changing the elevation

TABLE 4 | Detailed result of frequency of error (front).

Error of the estimated position	Number of epochs	Frequency	Cumulative relative frequency
1 m	923	0.22	0.22
2 m	1,676	0.40	0.63
3 m	809	0.19	0.82
4 m	268	0.06	0.89
5 m	139	0.03	0.92
6 m	77	0.02	0.94
7 m	46	0.01	0.95
8 m	32	0.01	0.96
9 m	32	0.01	0.96
10 m	147	0.04	1.00

TABLE 5 | Results of evaluation of the method using mileage and distance between positioning points.

Judgement	Number of epochs
(1) Result of the non-shadowed environment	
Judge: "small error"; real error: "small error"	293
Judge: "large error"; real error: "large error"	1
Judge: "large error"; real error: "small error"	3
Judge: "small error"; real error: "large error"	4
(2) Result of the shadowed environment	
Judgement: "small error"; real error: "small error"	127
Judgement: "large error"; real error: "large error"	70
Judgement: "large error"; real error: "small error"	40
Judgement: "small error"; real error: "large error"	36

masks using the result. Consequently, we confirmed that the error can be improved, albeit insignificantly.

7. POSSIBLE ERROR MAP APPLICATIONS

We discuss the possibility of the error map application in this section.

7.1. A-GPS

Assisted GPS (A-GPS) is a technology that uses information from base stations for positioning with mobile phones. A-GPS is useful for roughly grasping the user position from the received signal base station and narrowing down the user position. A technology called place engine performs positioning using Wi-Fi access points, even in indoor positioning, where satellite signals cannot be received. Therefore, in a shadowed environment, the positioning accuracy can be expected to improve by installation (e.g., beacons). For this purpose, a database of location information must be developed as an infrastructure. An error map can be used as a guide for setting up such infrastructure.

7.2. Elevation Mask Adjustment

We discussed this technique in the previous section. Satellites with low elevation angles generally often transmit low-quality signals. A receiver usually has a function that eliminates satellites transmitting low-quality signals by referring to the elevation angle. If the error map is used to estimate the "small error," the result is used to apply a weak elevation mask. Consequently, the overall accuracy of the positioning could be increased.

7.3. RTK

RTK is one of the technologies that has been drawing attention in recent years. It is a method of obtaining a mobile station's coordinates by finding the relative position with the mobile station on the user side by employing a reference station whose coordinates are known. The method accuracy is said to be of the order of a few centimeters, and high positioning accuracy can be achieved. The RTK system is required to count the number of satellite carrier waves at the user receiver. However, they may be interrupted by blocking the signal at a shadow, such as a building. Consequently, positioning may not be stable, even if the RTK is used in a shadowed environment. The degradation of the RTK positioning accuracy can be prevented by eliminating a satellite with a low-quality signal. It is possible to determine whether the environment is shadowed or not by using an error map; thus, it is possible to improve the positioning accuracy by using the variable elevation mask described to eliminate satellites that may cause signal interruptions.

7.4. Dynamic Map

Current positioning systems often use sensors to perform highly accurate positioning. Positioning support using dynamic maps for transportation is expected as a support method. Technologies that use map information, such as map matching, are often used. A system that supports each user's positioning should be put to practical use by employing the map information and the 3D building information, and smartphone and automobile information as big data. In this study, we examined the method for making a GNSS positioning error map as part of such a dynamic map. Furthermore, using 5G is expected to increase the communication speed, reduce the hardware size of IoT systems, and improve the data mining technology to enable an efficient use of big data. We expect the availability of dynamic maps in the future.

8. CONCLUSION

In this paper, we mentioned a cognitive navigation system inspired by the CogInfoCom that extends human cognitive capabilities and would even enable life support. Also, we described details of the error map as the technical basis of a cognitive navigation system.

An error map was designed to display the accuracy of the positioning of GNSS. We mentioned the method to measure the positioning error of GNSS by using two GNSS receivers. We designed the technique in two ways using two GNSS receivers, namely the estimation of error using two GNSS receivers and the comparison of the error size assisted by the error map.

First, we explained the problems related to positioning using GNSS and proposed the new error map idea to adopt the multipath environment and acquire a better position result. Next, we investigated the positioning error caused by the multipath under a shadowed environment when moving and stopping. Consequently, we confirmed that the distance of the two GNSS receivers and the average positioning error from these two receivers positively correlate. We proposed the two following methods according to the results: (1) error estimation by the linear regression; and (2) comparison of the error size assisted by the error map. We then evaluated the performance of the proposed method.

It is difficult to estimate the level of positioning error of GNSS. Usually, the number of satellites or Dilution of Precision (DOP) is used for that purpose. However, such methods cannot estimate the level of the error caused by multipath by buildings. A new idea to estimate the level of positioning error of GNSS is required by reflecting real filed positioning data. The proposed idea provides a useful and straightforward method to estimate the level of positioning error of GNSS by using a known distance of two GNSS receives.

Our method satisfied the target (error ≤ 1 m) of 20% of test data, but could not satisfy the target we set in sections 4 and 5. However, we can use our method for LBS, which does not require a critical boundary of accuracy. For the error size comparison, the no-shadow case provided an accuracy of over 90%; however, the accuracy was lower than 72% in the shadow case. We can distinguish the positioning stability; thus, we can offer a service to provide information about the possible error by using the error map and develop LBSs by incorporating the location's possible error.

The development of a real error map requires further study. If we develop an error map, we should consider several things,

including the useful number of error levels (such as 3 or 5 or 10) and the number of epochs for error calculation. Also, the idea of developing and maintaining the error map is required. One possibility is to attach GNSS receivers on buses, trucks, and taxis to collect the position data and errors. The map must be frequently updated after developing the error map. This map provides information about the error level, and the user can rely on it as information to support positioning (e.g., shadow mask application to obtain more accurate position data or understand the position data reliability).

We think that many fields can be extended usability and performance by using the idea of CogInfoCom. We would like to try to use the concept of CogInfoCom in the different research areas.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

The part of the research results was obtained from the commissioned research by the National Institute of Information and Communications Technology (NICT), JAPAN. Also, this research was supported by JSPS KAKENHI Grant Number JP17H02249, JP18K11849, and JP17H01731.

REFERENCES

- Baranyi, P., and Csapo, A. (2012). Definition and synergies of cognitive infocommunications. *Acta Polytech. Hung.* 9, 67–83.
- Baranyi, P., Csapo, A., and Sallai, G. (2015). *Cognitive Infocommunications (CogInfoCom)*. Cham: Springer International.
- Iwase, T., Suzuki, N., and Watanabe, Y. (2013). Estimation and exclusion of multipath range error for robust positioning. *GPS Solut.* 17, 53–62. doi: 10.1007/s10291-012-0260-1
- Kitani, T., Hatano, H., Onishi, H., and Hirao, S. (2012). *A Method to Improve Positioning Accuracy of GPS by Sharing Error Information Among Neighbor Devices*. Information Processing Society of Japan SIG Technical Report in Japanese 2012-ITS-49, 1–6.
- Klobuchar, J. A. (1987). Ionospheric time-delay algorithm for single frequency GPS users. *IEEE Trans. Aerospace Electr. Syst.* 23, 325–331. doi: 10.1109/TAES.1987.310829
- Kubo, N., Kobayashi, K., Hsu, L.-T., and Amai, O. (2017). Multipath mitigation technique under strong multipath environment using multiple antennas. *J. Aeronaut.* 49, 75–82.
- Misra, P., and Enge, P. (2001). *Global Positioning System Signals, Measurements, and Performance*. Lincoln, MA: Ganga-Jamuna Press.
- Miyata, H., Noguchi, T., Sakitani, A., and Egashira, S. (1996). Consideration on high accurate positioning with simplified DGPS. *ITE Tech. Rep. (in Japanese)* 20, 71–76.
- Miyata, H., and Sakitani, A. (1997). Improving accuracy of simplified dgps positioning by the after processing. *ITE Tech. Rep. (in Japanese)* 21, 19–22.
- Nakamura, Y. (2018). *A research on reduction method of positioning error based on error characteristics under the same combination of gnss satellites*. Utsunomiya: Utsunomiya University Graduation thesis (in Japanese).
- Odaka, Y., Takano, S., In, Y., Higuchi, M., and Murakami, H. (2011). “The evaluation of the error characteristics of multiple GPS terminals,” in *Proceedings of the 2nd International Conference on Circuits, Systems, Control, Signals (CSCS '11)* (Prague), 13–21.
- Osechas, O., Kim, K. J., Parsons, K., and Sahinoglu, Z. (2015). “Detecting multipath errors in terrestrial GNSS applications,” in *Proceedings of the 2015 International Technical Meeting of The Institute of Navigation* (Dana Point, CA), 465–474.
- Rohani, M., Gingras, D., and Gruyer, D. (2016). A novel approach for improved vehicular positioning using cooperative map matching and dynamic base station DGPS concept. *IEEE Trans. Intell. Transport. Syst.* 17, 230–239. doi: 10.1109/TITS.2015.2465141
- Saastamoinen, J. (1973a). Contributions to the theory of atmospheric refraction. *Bull. Geodesique* 30, 279–298.
- Saastamoinen, J. (1973b). Contributions to the theory of atmospheric refraction. *Bull. Geodesique* 30, 383–397.
- Saastamoinen, J. (1973c). Contributions to the theory of atmospheric refraction. *Bull. Geodesique* 30, 13–34.
- Sakai, T. (2003). *Introduction of GPS*. Tokyo: Tokyo Denki University Press.

- Satirapod, C., and Chalermwattanachai, P. (2005). Impact of different tropospheric models on gps baseline accuracy: case study in Thailand. *J. Glob. Posit. Syst.* 4, 36–40. doi: 10.5081/jgps.4.1.36
- Vincenty, T. (2013) (1975 (Published online: 2013)). Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Surv. Rev.* XXIII, 88–93. doi: 10.1179/sre.1975.23.176.88
- Watanabe, Y., Sato, K., and Takada, H. (2020). Dynamicmap 2.0: a traffic data management platform leveraging clouds, edges and embedded systems. *Int. J. Intell. Transport. Syst. Res.* 18, 77–89. doi: 10.1007/s13177-018-0173-7

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.

Copyright © 2021 Ito, Nakamura, Hiramatsu, Kitani and Hatano. 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.



Speech Pauses and Pronominal Anaphors

Costanza Navarretta *

Centre for Language Technology, Department of Nordic Studies and Linguistics, University of Copenhagen, Copenhagen, Denmark

OPEN ACCESS

Edited by:

*Anna Esposito,
University of Campania Luigi Vanvitelli,
Italy*

Reviewed by:

*Rubén San-Segundo,
Polytechnic University of Madrid,
Spain
Dubravko Culibrk,
University of Novi Sad, Serbia*

***Correspondence:**

*Costanza Navarretta
costanza@hum.ku.dk*

Specialty section:

*This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science*

Received: 27 January 2021

Accepted: 28 April 2021

Published: 13 May 2021

Citation:

*Navarretta C (2021) Speech Pauses
and Pronominal Anaphors.
Front. Comput. Sci. 3:659539.
doi: 10.3389/fcomp.2021.659539*

This paper addresses the usefulness of speech pauses for determining whether third person neuter gender singular pronouns refer to individual or abstract entities in Danish spoken language. The annotations of dyadic map task dialogues and spontaneous first encounters are analyzed and used in machine learning experiments act to automatically identify the anaphoric functions of pronouns and the type of abstract reference. The analysis of the data shows that abstract reference is more often performed by marked (stressed or demonstrative pronouns) than by unmarked personal pronouns in Danish speech as in English, and therefore previous studies of abstract reference in the former language are corrected. The data also show that silent and filled pauses precede significantly more often third person singular neuter gender pronouns when they refer to abstract entities than when they refer to individual entities. Since abstract entities are not the most salient ones and referring to them is cognitively more hard than referring to individual entities, pauses signal this complex processes. This is in line with perception studies, which connect pauses with the expression of abstract or complex concepts. We also found that unmarked pronouns referring to an entity type usually referred to by a marked pronoun are significantly more often preceded by a speech pause than marked pronouns with the same referent type. This indicates that speech pauses can also signal that the referent of a pronoun of a certain type is not the most expected one. Finally, language models were produced from the annotated map task and first encounter dialogues in order to train machine learning experiments to predict the function of third person neuter gender singular pronouns as a first step toward the identification of the anaphoric antecedents. The language models from the map task dialogues were also used for training classifiers to determine the referent type (speech act, event, fact or proposition) of abstract anaphors. In all cases, the best results were obtained by a multilayer perceptron with an F1-score between 0.52 and 0.67 for the three-class function prediction task and of 0.73 for the referential type prediction.

Keywords: speech pauses, abstract pronominal anaphora, individual pronominal anaphora, machine learning, corpus annotation, map task dialogues, first encounters

1 INTRODUCTION

This paper addresses the role of speech pauses for determining what third person neuter gender singular pronouns, such as the English *it*, *this* and *that* refer to. The pronouns we address are pronominal intersentential anaphors, which can refer to concrete or abstract entities depending on their context. In example 1, the pronoun *it* refers to the glass on the table in the given communicative setting and has the nominal phrase *the glass* as antecedent. Instead in example 2, the pronoun refers to the generic event of cleaning the machines before and after use, and points back to the clause *Clean the machines before and after use*.

1. The glass is on the table. Please bring *it* to me. (antecedent: the glass).
2. Clean the machines before and after use. Always do *it* carefully. (antecedent: clean the machines before and after use.)

The other abstract pronouns in English are the demonstrative *this* and *that*, which often refer to complex abstract entities of different type. Pronominal anaphors that refer to individual entities as in 1 are called individual anaphors, while they are called abstract anaphors, discourse deictics (Webber, 1988), or situational referents (Fraurud, 1992) if they refer to an abstract entity as in 2. Third person neuter gender singular pronouns, *tpngs* pronouns henceforth, are quite common in all languages. Determining what they refer to, especially in the case of abstract anaphors, can be difficult for humans, and it is one of the most complex tasks in natural language processing. The task of finding the referent of referring expressions is called coreference resolution. In most natural language processing (NLP) systems, the task is simplified and resolution is redefined as finding the immediate *antecedents* of referring expressions. This process results in the annotation of co-referential chains of varying length as well as of expressions that are only mentioned once in discourse, the so called singletons. Finding the antecedents of intersentential pronominal anaphors is a sub-task of coreference resolution, called pronominal anaphora resolution. Most coreference systems only address expressions referring to individual entities, but more recently also events are dealt with e.g., Yang and Mitchell (2016), Barhom et al. (2019).

In this paper, we investigate the phenomenon of abstract and individual *tpngs* anaphors in Danish dialogues focusing on the importance of stress (accent) information on words and of speech pauses preceding the anaphors in order to determine whether they refer to individual or abstract entities or have other functions. Stress on words has especially been studied in terms of information structure, because words that are marked by a main accent are emphasized by the speaker and therefore become more salient to the addressee. Since salience is one of the main factors influencing the interpretation of pronominal anaphors, stress information is relevant in our context. Moreover, speech pauses have been interpreted as cognitive signals showing that the speaker is going to produce an abstract or complex concept Rochester (1973), Reynolds and Paivio (1968) or is going to

refer to an object who has not the highest salience level e.g., (Gargiulo et al., 2019). More precisely, we analyze the relation between stress, speech pauses and *tpngs* pronouns in Danish annotated dialogues and then use this relation in language models on which we train classifiers to discriminate individual and abstract anaphors from other uses of these pronouns, and to predict the referent type of abstract anaphors. For the first classification task, the annotations of map task and first encounters dialogues are used, while for the second task we only use the annotations of the map task dialogues.

In **section 2**, studies about individual and abstract pronominal anaphora are presented, and the Danish *tpngs* pronouns are introduced. In **section 3**, relevant research on speech pauses is discussed. In the following **section 4**, the data used in this work is described. Moreover, the occurrences of abstract and individual *tpngs* pronouns in the two dialogue types and the role of stress and speech pauses for discriminating their uses are analyzed. In **section 5**, machine learning experiments are presented in which classifiers are trained on language models consisting of words alone, or enriched with accent information and speech pauses, in order to identify the pronouns' function and referent type. Finally, we discuss the results of our study in **section 6** and conclude in **section 7**.

2 INDIVIDUAL AND ABSTRACT PRONOMINAL REFERENCE

When we talk or write, we refer continuously to entities through nominal phrases. These can be substantives, with or without determiners and modifiers, or personal and demonstrative pronouns. Since *tpngs* pronouns can refer to either individual or abstract entities, it is first necessary to determine whether the referent is abstract or concrete in order to find their antecedent.

According to all researchers, who have presented cognitive models that account for the choice of referring expression by a speaker, this choice is based on assumptions the speaker makes about the varying status of entities in the addressee's mental state, see inter alia Prince (1981), Prince (1992), Gundel et al. (1993), Givón (1979), Ariel (1988), Ariel (2001). In all models, pronouns refer to the most salient or accessible entities. However, salience or accessibility is defined differently in the various models, even though all definitions are related. Prince (1981) and Prince (1992) looks at information structure and considers the most salient entities those that are oldest or known for longest time in discourse. Givón (1979) proposes to consider the entities that are most *topic prominent* at that point in discourse as being the most salient ones. According to (Gundel et al., 1993) instead, the most salient entities are those that are *in focus*. They also propose a scale of salient referring expressions, the *Givenness Hierarchy*. Finally, Ariel (2001) introduces the concept of *accessibility* level of entities, and she operationalizes it in terms of the distance of the referring expressions from their antecedents. In Ariel's model, the least marked and therefore easiest accessible referring expressions are zero pronouns, while the most marked expressions are full names with modifiers. Ariel's accessibility scale also distinguishes

between unstressed and stressed pronouns, the first being more accessible than stressed pronouns.

Kameyama (1998) studies the occurrences of unstressed and stressed versions of the same pronouns when they occur in the same position in discourse. She concludes that they have the same denotational range but indicate different preferred values. Because stressed pronouns signal a different presupposition than their unstressed counterparts, stressed pronouns take the complementary preference of their unstressed equivalents, when there are competing antecedents. Kameyama implements these findings in a system of preferences added to the Centering framework.

Webber (1988) finds that the English personal pronouns *it* does not often refer to abstract entities when the antecedents are clauses, because clauses are not easily accessible in discourse. The demonstrative pronouns *this* and *that* are instead used in these cases. Gundel et al. (1993) find the same pattern in their data and explain the preferred use of demonstrative pronouns with clausal antecedents in terms of their Givenness Hierarchy. Entities introduced in discourse by clauses are only activated in the hearer's cognitive status, being their referents facts, situation and propositions. On the contrary, eventualities, which are introduced in discourse by verbal or nominal phrases, often occur in a central syntactic position in the current or in the preceding utterance, and they are therefore often *in focus* and can be referred to by personal pronouns.

The preference for using demonstrative pronouns when referring to clauses or utterances as well as to discourse segments has been implemented in few rule-based resolution systems for English e.g., Eckert and Strube (2001), Byron (2002), Strube and Müller (2003), Müller (2007). An adaptation of Eckert and Strube (2001)'s algorithm to Danish was presented in Navarretta (2002) and Navarretta (2004).

The corpora annotated with abstract pronominal anaphora are only few (Kolhatkar et al., 2018), comprising the ARRAU corpus, which consists of both texts and transcriptions of dialogues (Poesio and Artstein, 2008). Part of the ARRAU corpus, the Wall Street Journal (WSJ) texts, has been used in the first benchmark for the resolution of *tpngs* anaphors in English (Marasovic et al., 2017). More precisely, Marasovic et al. (2017) run an LSTM-Siamese Net on the ARRAU annotations and report a precision score (s@1 which indicates that the correct antecedent span has been marked) of 29.06. Moreover, Poesio et al. (2018) describe how the ARRAU corpus was prepared to find the antecedents of abstract pronominal anaphors for a shared competition (CRAC, 2018), but no research group has addressed this task. The most recent work in which coreference resolution includes abstract pronouns is (Uryupina et al., 2020). The authors report an F1-score of 49.18 for the identification of non anaphoric expressions and singletons achieved by a LSTM. All morphological and syntactic features annotated in the data were included as well as all types of referring expressions in the ARRAU corpus.

Not only it is difficult to find the correct antecedents of abstract anaphors as shown in (Marasovic et al., 2017), it is also difficult to decide what the referent is (a task not addressed in anaphora resolution), because the anaphors often create their

referent in the moment they point back to an antecedent Webber (1991), Fraurud (1992). Examples of how the same clausal antecedent of an abstract anaphor has different referents, depending on the context in which the anaphor is uttered, are in 3a–3c.

3. Peter was injured in a car accident
 - a. When did *that* happen? (an event)
 - b. *That* is not true. (a proposition)
 - c. I did not know *that*. (a fact)

The three occurrences of the demonstrative *that* in the three utterances 3a–3c have the same antecedent, the utterance 3 *Peter was injured in a car accident*. However, in the example 3a, the referent is an event, in 3b it is a proposition and in 3c is a fact. Fraurud (1992) proposes that all abstract situations, defined by Vendler (1963) can be the referents of abstract anaphors. Asher (1993) builds a hierarchy of what he calls saturated abstract objects ordering abstract entities with respect to their degree of abstractness. The less abstract entities are eventualities, while the most abstract ones are proposition-like entities. Fact-like entities are in the middle. Also according to Asher, the entities which have the lowest degree of abstractness, can be referred to by personal pronouns as also noticed by Webber (1991), Gundel et al. (1993), while the most abstract entities are referred to by demonstrative pronouns.

The most common Danish *tpngs* pronoun is *det*, which corresponds to the three English pronouns *it*, *this* and *that*. In written language, *det* is ambiguous since it is not possible to determine whether it is a personal or a demonstrative pronoun. In spoken language, the unstressed *det* is the personal pronoun, while its stressed version *d,et*¹ is a demonstrative pronoun corresponding to both the English *this* and *that*. The demonstrative form can also occur as two words *det her* (this) and *det der* (that)². Another Danish *tpngs* demonstrative pronoun is *dette* (this), which is only rarely used in spoken language.

The Danish *det* has many functions, and when it is used as anaphor, it can either refer to an individual or an abstract entity. In the first case, it can have as antecedents a nominal phrase referring to concrete entities in all genders and numbers, while in the second case it has other types of syntactic phrases as antecedents: verbal phrases, predicates in copula constructions, clauses and discourse segments (Navarretta, 2002). It has been found that Danish *tpngs* abstract pronouns occur in different contexts than the corresponding English pronouns and therefore theories and algorithms accounting for the use of these English pronouns cannot be directly applied to Danish data Navarretta (2002), Navarretta (2004). A Danish corpus annotated with abstract and individual *tpngs* anaphora was produced in the DAD project Navarretta and Olsen (2008). Our current study uses part of this corpus as well as the annotations of eight dyadic

¹In this article, stress is marked with a comma preceding the stressed vowel.

²Allan et al. (1995) call the stressed versions of the pronoun *det* for *emphatic* pronouns.

first encounters from the Danish NOMCO corpus (Paggio and Navarretta, 2017) coded for this study.

3 RELEVANT STUDIES ON SPEECH PAUSES

Different types of speech pauses have been identified in discourse, the main types being silent pauses and filled pauses. Filled pauses are pauses accompanied by audible breaths or sighs, but can also occur with more or less lexicalized items, such as the English *ah* and *uh*. These items have been called fillers, discourse particle, and discourse markers.

Researchers have identified multiple functions of pauses involving both speech production and perception. For example, pauses have been found to be indicators that the speaker is planning what (s) he is going to say (MacLay and Osgood (1959), Goldman-Eisler (1968), Chafe (1987), Hirschberg and Nakatani (1998), Shriberg (1994), and, they can limit syntactic phrases. Pauses can also signal that the speaker is searching for the correct lexical item (Krauss et al., 2000) and they are one of the multi-modal signals that contribute to regulate turn taking (Duncan and Fiske (1977), Allwood (1988), Clark and Fox-Tree (2002).

Reynolds and Paivio (1968) found that silent and filled pauses occurred more frequently when students had to provide definitions of abstract objects than when they had to explain concrete objects. In line with this study, Rochester (1973) investigated the frequency of hesitations in discourse and he concluded that their occurrences increase when speakers have to express something difficult or have to choose between more options. Finally, Esposito et al. (2002); Esposito and Esposito (2011) studied the occurrences of pauses and gestural pauses in more languages and propose that some of them have the function of introducing discourse new information either by speech or by gestures.

Navarretta (2007) analyzed the relation between types of Danish and Italian *tpngs* pronouns and their functions in an annotated abstract anaphora corpus, the DAD corpus (Navarretta and Olsen, 2008). She trained a support vector machine on the Danish annotations of texts, monologues and dialogues in order to identify these functions automatically (Navarretta, 2010) analyses the experiments focusing on the role of stress information in the spoken part of the data. Navarretta (2007)'s analysis of the data showed that there are numerous abstract and concrete pronominal occurrences in the data, and that both stressed and unstressed pronouns refer frequently to individual and abstract anaphors, confirming preceding studies that indicated that personal pronouns refer to abstract entities in Danish much more frequently than in English (Navarretta, 2004). In the classification experiments, run in WEKA, language models consisting of unigrams, bigrams and trigrams preceding and following the *tpngs* pronouns were used as training data and the best results were obtained when stress information was used and silent pauses were added to the language models consisting of different types of n-grams. The best results when classifying the functions of pronouns gave an F1

score of 0.51. The monologues contained only very few silent pauses, and the results of classification were extremely high (F1 score 0.982). These high results are due to the fact that the monologues consisted of the same texts read up by the various participants, and therefore train and test data were nearly the same with only small differences determined mostly by hesitations and placement of the stress by the different readers.

In a manually annotated spoken corpus, Roesiger and Riester (2015) find that information about the accent on words can contribute to coreference resolution. Moreover, Roesiger et al. (2017) test these findings on German data and conclude that pitch accents and phrasing improve coreference resolution. Gargiulo et al. (2019) study prosodic features and overt pronouns with individual antecedents in subject or object position in a production study involving Italian and Swedish speakers. They conclude that in both languages longer pauses precede inter-clausal pronouns when the antecedent is the less expected one. This is the subject for an overt pronoun in Italian, and the object in Swedish. Therefore, the function of pauses proposed by Gargiulo et al. (2019) is similar to the function of stress on the pronouns described by Kameyama (1998).

In this paper, we build upon the work in Navarretta (2007), Navarretta (2010) and re-use part of the data from those studies, as well as the newly annotated first encounters dialogues. The preceding studies are therefore extended in the following way: 1) both silent and filled pauses are considered in the present work, 2) a statistical analysis of differences of occurrences between individual and abstract pronouns is performed on the two types of dialogues, 3) anaphoric pronouns are studied together with the pauses which precede them in the maptask dialogues in order to determine whether they have specific uses, while Navarretta (2010) only added pauses to the training data as an extra prosodic feature without analyzing their possible functions, 4) we perform a number of classification experiments aimed to identify automatically three functions of *tpngs* Danish pronouns (individual, abstract entities or other functions) in the map task and first encounters dialogues, 5) we apply classifiers to the map task data in order to train classifiers to identify automatically the referent type of the abstract anaphors with verbal and clausal antecedents. In all experiments, we focus especially on the effect of pauses and stress information on classification.

4 THE DATA

4.1 The Annotated Corpora

The data used in this study consist of part of the DAD corpus and part of the NOMCO corpus.

The DAD corpus was collected and annotated under the project DAD, Det Abstrakt Det (the abstract it), funded by the Danish Research Councils. It consists of a collection of Danish and Italian texts and spoken data that were annotated with information about the antecedents and referents of *tpngs* pronouns. The annotations were made by two expert annotators (Navarretta and Olsen, 2008). In the present study, we only use the part of the DAD corpus consisting of the DanPASS dialogues (Grønnum, 2009). The DanPASS

dialogues are a Danish version of the map task dialogues described in (Anderson et al., 1991). In the map task dialogues, one participant guided the second participant in going through a map from the start point to an end point, the target. The two participants were sitting in two different rooms and could only speak together through head sets with microphones. The task was made more complex by the maps that the two participants worked with. The maps were similar, but they also had some differences, and the participants were not told about this.

The DanPASS corpus was collected, transcribed and phonetically annotated by phoneticians at the University of Copenhagen. The corpus consists of read texts, and map task dialogues. The participants were students and employees from the Department of General and Applied Linguistics, today part of the Department of Nordic Studies and Linguistics, at the same university. Stress, hesitations, pauses, tone of voice and other phonetic features were manually annotated by phoneticians (Grønnum, 2009) using PRAAT (Boersma and Weenink, 2009). The running words in the dialogue transcriptions are 52,145. In the DAD project, the transcriptions with pause and stress information were converted to XML and *tpngs* pronominal anaphora information were added using the PALinkA tool (Orăsan, 2003). The annotation scheme followed for coding anaphoric information was an extension of the MATE/GNOME annotation scheme for anaphora (Poesio, 2004). The extensions consisted mainly of elements and attributes added to the scheme in order to code non referential uses of the pronouns I and describe e.g., the abstract referent type. Most of the DAD data were annotated independently by the two annotators and then compared. Inter-coder agreement in terms of kappa scores Cohen (1960), Carletta (1996) was between 0.7 and 1 (Navarretta and Olsen, 2008), depending on the class and attribute. The data that were annotated by only one coder, was controlled by the second coder. In case of disagreement, the two annotators decided together which annotation to adopt. In difficult cases, linguist colleagues were consulted to choose the most probable annotation³.

The Danish NOMCO corpus of first encounters consists of twelve spontaneous audio- and video-recorded dialogues between two young people who meet for the first time and talk freely for about few minutes. The dialogues were collected and transcribed under the Nordic NOMCO project (Navarretta et al., 2012) at the University of Copenhagen. The transcriptions include pause and stress information annotated in the same format and system as in the DanPASS project, and the corpus has especially been used for studying multi-modal communication. Eight of the NOMCO encounters have been annotated with *tpngs* pronouns for this study. The *tpngs* pronoun annotations have been coded in an excel file following the annotation scheme and annotation manual produced by the DAD project. One dialogue was annotated by two coders independently, while the remaining

dialogues were annotated by one coder. The intercoder agreement obtained for the annotations performed by two annotators in terms of kappa scores is between 0.58 and 1 depending on the category. These scores are lower than that obtained on the DAD corpus, which covered both texts and map task dialogues. The total of tokens comprised in these annotations are 13,300 and the total duration of the eight dialogues is of approx. 40 min.

4.2 The Annotations

The pronominal uses annotated in the data are the following:

- pleonastic as in *det sner* (it snows), *hun har det godt* (lit. she has it fine) (She is fine);
- cataphoric, the pronoun precedes the linguistic expression that is necessary to interpret it. *Det at Hanne ikke blev færdig med analysekapitlet til tiden, skabte problemer for hendes medstuderende.* (lit. It that Hanne did not finish the analysis chapter in time, gave problems to her fellow students) (The fact that Hanne did not finish the analysis chapter in due time, gave problems to her fellow students);
- deictic. The pronoun refers to an object in the physical context as in the following example: *Hvad er det der?*
- (What is that?)
- possibly accompanied by a pointing gesture to an object;
- individual anaphoric, the antecedent is a concrete entity: A: *Jeg har det forladte kloster lige i midten af kortet.* B: *Ja, det har jeg sådant set også.* . . (A: I have the abandoned closter precisely in the middle of the map B: I have it also, in a way . . .)
- individual vague anaphoric⁴, the antecedent of the pronoun is a concrete entity that is implicit in discourse.
- abstract anaphoric, the antecedent is an abstract object: A: *okay der har jeg noget der hedder Den Blå Sø* B: *nej men det er jo helt forkert* (A: okay there I have something that is called The Blue Lake B: no but this is completely wrong)
- textual deictic (Lyons, 1977). “*Jeg elsker dig*” - *Det sagde han til hende for første gang, mens de snakkede.* (“I love you” - He said that to her for the first time, while they talked together);
- abstract vague anaphoric-the abstract antecedent is implicit in the discourse;
- abandoned: the pronoun occurs in an unfinished utterance, which is then abandoned, and therefore it is not possible to infer the referent:
- *det er - han er gået*
- (it is - He is gone)

The type of referent of the abstract anaphors was also annotated in the map task dialogues as one of the extensions to the MATE/GNOME annotation scheme. The referent types that were identified are *eventuality*, *fact-like*, *fact-event*, *proposition-like*, *speech-act*. The type *fact-event* was assigned when the annotators found that the referent of the abstract anaphor was ambiguous and could be either a fact or an event.

³In some cases, when the annotators recognized that an anaphor could be interpreted in various way with the same probability, a class covering both readings was proposed. This was the case for the classification of the referent types.

⁴The use of the term *vague* in the list is taken from Eckert and Strube (2001).

TABLE 1 | Frequencies of pronominal uses in DanPass dialogues.

Pronoun	Indiv	IndVag	Abstr	AbstVag	Pleon	Cathaphor	Deic	Textdeict	Aband	Total
Unmarked	176	25	100	5	44	17	0	4	103	434
Marked	123	22	110	7	0	22	7	3	0	334
Total	299	47	219	12	44	39	7	7	103	768

4.3 Data Analysis

There are 768 *tpngs* pronouns in the DanPASS dialogues, and the occurrences of each type of pronoun are the following:

- *det*: 433
- *d,et*: 322
- *det her*: 1
- *d,et her*: 2
- *det h,er*: 1
- *det d,er*: 1
- *d,et h,er*: 3
- *det d,er*: 2
- *d,et der*: 1
- *det der*: 1
- *dette*: 1

The only occurrence of the demonstrative *dette* is an individual anaphor. When analyzing the demonstrative pronouns *det her* and *det der* and *dette*, we will consider them to be marked as the pronoun *d,et* independently from the presence and/or position of the stress. In **Table 1**, we show the functions of personal (non stressed/unmarked *det*) and demonstrative (marked) pronouns, that is *d,et*, *dette*, *det der*, and *det der*.

As expected, most occurrences of the pronouns have an anaphoric function, and the most common anaphoric use is that of referring to an individual entity (345 occurrences), followed by reference to abstract entities (223). 201 individual anaphors are unmarked, and 144 are marked. 105 abstract anaphors are unmarked, and 118 are marked. The difference between the use of unmarked and marked pronouns in reference to individual vs. abstract entities is statistically significant. The chi-square is 6.8076, the p-value is 0.009077, with $df = 1$. The result is significant in the confidence interval $0.92 > p < 0.008$.

Thus, also in Danish spoken language there is a significant preference for referring to individual entities through unmarked pronouns and to abstract entities through marked ones. Textual deixis is performed by both marked and unmarked pronouns, and there are more marked cataphors than unmarked ones. Finally, abandoned pronouns are always unmarked in the DanPASS dialogues.

Out of the 345 occurrences of individual anaphors, 80 are preceded by a silent pause and 6 are preceded by a filled pause, that is 25% of the individual anaphors are preceded by a pause. 76 out of the 223 abstract anaphors are preceded by a silent pause and 9 of them by a filled pause, for a common total of 38% of their occurrences. Also in this case, the difference between the two types of anaphora is statistically significant. More precisely, considering the two types of pause together, the chi square

TABLE 2 | Abstract referent types and pronominal types in the DanPass dialogues.

Referent type	Pause + Unmark	Unmark	Pause + Marked	Marked	Total
Eventuality	7	19	9	31	66
Fact-event	1	1	1	0	3
Fact	15	13	17	32	77
Proposition	17	22	8	11	58
Speech-act	3	1	1	2	7
Total	43	56	36	76	211

statistic is 11.1973. The p-value is 0.000819 with degree of freedom = 1. The result is significant at $0.98 > p < 0.02$. Considering each type of pause separately, the chi square statistic is 11.9277. The p-value is 0.007637, with $df = 3$. The result in this case is significant in the confidence interval $0.0002 > p < 0.9998$. This shows that pauses preceding *tpngs* pronominal anaphors more frequently signal that the speaker is going to refer to an abstract entity than to an individual entity. As we have discussed previously, abstract reference is more complex and less expected than individual reference.

In **Table 2**, the referential types of the various abstract pronouns in the DanPASS data are shown, distinguishing those that are preceded by pauses and those that are not.

The table shows that the most frequent referent type of abstract anaphors in these data is a fact-like entity, followed by eventualities and propositions. Surprisingly eventualities are more often referred to by a marked pronoun than by an unmarked pronoun, while propositions are more often referred to by unmarked pronouns than by marked ones. This seems to be contrary to what found for English by e.g., Webber (1988), Gundel et al. (1993) who notice that eventualities can be more easily referred by the personal pronoun *it* than other abstract entities since events are the most accessible abstract entities in discourse. It is interesting that there is a preference for pauses to precede unmarked pronouns more frequently when they refer to propositions and facts than when they precede eventualities. In this case, the chi-square statistic is 3.3403. The p-value is 0.067602 with $df = 1$. The confidence interval is $0.07 > p < 0.093$. This can be interpreted as a signal that pauses which precede a personal abstract pronoun in some cases signal that the referent of the anaphor is of a less expected type given its referent. This could be a similar function to that identified for other referential phenomena by (Gargiulo et al., 2019). However, the frequency of unmarked pronouns as referent of entities of higher abstractness degree also confirms the observation that there are language specific differences with respect to individual and

TABLE 3 | Frequencies of pronominal uses in the NOMCO encounters.

Pronoun	Indiv	IndVag	Abstr	AbstVag	Pleon	Cathaphor	Deic	Textdeict	Aband	Total
Unmarked	172	16	165	14	97	75	0	2	54	594
Marked	22	6	61	8	0	8	12	3	22	142
Total	194	21	226	22	97	81	12	5	76	736

abstract reference between Danish and English as discussed in Navarretta (2002), Navarretta (2004).

There are 736 *tpngs* pronouns in the annotated first encounters. The pronouns are distributed into the following types:

- *det*: 594
- *d,et*: 127
- *det her*: 1
- *d,et her*: 1
- *det der*: 4
- *det d'er*: 1
- *d,et der*: 7
- *d,et d,er*: 1

There are 594 unmarked and 142 *tpngs* pronouns in these dialogues. It is interesting to note that *tpngs* pronouns are much more frequent in the spontaneous first encounters than in the map task dialogues, since their relative frequency in the former corpus is 0.055, while it is 0.015 in the latter.

In **Table 3** are shown the different types of pronominal functions of the *tpngs* pronouns in the first encounters. Unmarked *tpngs* pronouns are more frequent in these dialogues than in the DanPASS dialogues, while individual *tpngs* anaphors are more frequent in the map task dialogues than in the first encounters. Finally, abstract *tpngs* anaphors are more frequent in the first encounters than in the map task dialogues.

The fact that individual anaphors are more frequent in the map task dialogues is not surprising given that the speakers often refer to individual objects on their maps. The higher frequency of abstract anaphors in the first encounters can explain the lower intercoder agreement obtained for these data, since individual anaphors are easier to annotate. In the first encounters, the unmarked pronoun is more often used as individual than as abstract anaphor, while the opposite holds for marked pronouns. The difference is also in these dialogues significant. In fact, the chi-square statistic is 15.417. The p-value is 0.000086 with $df = 1$ and the confidence interval holds for $0.0001 < p < 0.9999$. This difference confirms again that also in spoken Danish as in English in general there is a preference for unmarked pronouns to refer to individual objects, and for marked pronouns to refer to abstract pronouns. However, these data also confirm that unmarked pronouns are also used as abstract anaphors in Danish more frequently than in English. In fact, the pronoun *it* has been found to be abstract pronoun in less than one third of its occurrences in e.g., Webber (1988), Gundel et al. (1993), Poesio and Artstein (2008). The analysis of pauses preceding *tpngs* pronoun types will be performed in future.

5 IDENTIFYING PRONOMINAL ANAPHORIC FUNCTIONS AND ABSTRACT REFERENT TYPES

In our classification experiments, we address the automatic classification of the individual and abstract functions of *tpngs* pronouns in the Danish map task and first encounters dialogues. Differing from the experiments presented by Navarretta (2007), Navarretta (2010), we include filled pauses to the annotations and focus on the classification of individual and abstract anaphors vs. all other uses of the pronouns. Moreover, a different classification strategy and more classifiers are applied.

In these experiments, the five categories cataphoric, pleonastic, deictic, textual deictic and abandoned have been collapsed in one class *other*, since our main aims are 1) to determine whether the information of stress and pauses can help disambiguating individual and abstract anaphors as well as distinguish them from other uses, and 2) to find the best data sets and classifiers for this task.

The DanPASS data sets contain 345 pronouns referring to an individual entity (the individual and individual vague pronouns), 223 pronouns referring to an abstract entity (the abstract and abstract vague pronouns), and 200 pronouns classified as *other*. The NOMCO data sets contain 216 individual pronouns (also in this case the pronouns classified as individual and individual vague), 248 abstract pronouns (those classified as abstract and abstract vague) and 272 pronouns classified as *other*. The supervised machine learning experiments were run in python 3.7 with the numpy and scikit-learn packages. The classifiers that were tested are K Neighbors Classifier (KNC), Multinomial Naive Bayes (MNB), Multilayer Perceptron (MP), Support Vector Machine (SVM), and Logistic Regression (LR).

The multilayer perceptron was run with the following hyper-parameters: the *adam* solver with the *tahn* activation, two layers of size 3 and 1, 5,000 iterations, adaptive learning rate and $\alpha = 0.001$ for the DanPASS data. The parameters used for the NOMCO corpus are the *sgd* solver with the *tahn* activation, four layers of sizes 8, 5, 5, and 1, constant learning rate and $\alpha = 0.001$. The optimal parameters were found applying the GridSearchcv method on a variety of hyper-parameters including two solvers, two learning rates and activation layers as well as different number of layers. The GridSearchcv was run on 20% of the data, then a different 20% of the data was used for testing, and finally, ten-fold cross-validation for the best performing classifier and data set were run to control that the results from testing are not due to overfitting.

The language models in the training data were the ones that gave the best results on the task of classifying all pronominal functions in (Navarretta, 2007). The models consist of six-grams,

TABLE 4 | Examples of various data sets.

Data set	Token1	Token2	Token3	Token4	Token5	Token6
Word	jeg	men	det	er	jo	Ikke
word&stress	jACCeg	men	det	er	jo	iACCkke
word&pause	FILPAUSE	men	det	er	jo	Ikke
All	FILPAUSE	men	Det	er	jo	iACCkke

TABLE 5 | Results of classifiers predicting pronominal type in the DanPASS dialogues.

Classifier	Dataset	P	R	F1
Majority		0.152	0.39	0.22
Random		0.322	0.312	0.314
KNC	Words	0.546	0.558	0.541
	words&accent	0.57	0.571	0.572
	words&pause	0.514	0.523	0.51
	All	0.585	0.591	0.582
	Word	0.647	0.623	0.62
MNB	word&accent	0.683	0.81	0.675
	word&pause	0.664	0.649	0.649
	All	0.698	0.681	0.682
	Words	0.738	0.662	0.651
SVM	word&accent	0.7	0.636	0.636
	word&pause	0.714	0.662	0.628
	All	0.71	0.623	0.6
	Word	0.638	0.617	0.619
MP	word&accent	0.645	0.643	0.643
	word&pause	0.65	0.63	0.64
	All	0.69	0.681	0.684
	Words	0.663	0.643	0.64
LR	word&accent	0.647	0.636	0.637
	word&pause	0.67	0.662	0.663
	All	0.685	0.67	0.67

TABLE 6 | Results of classifiers predicting pronominal type in NOMCO dialogues.

Classifier	Dataset	P	R	F1
Majority		0.114	0.339	0.171
Random		0.333	0.331	0.331
KNC	Words	0.466	0.466	0.466
	words&accent	0.46	0.46	0.46
	words&pause	0.393	0.4	0.395
	All	0.44	0.44	0.44
MNB	Word	0.523	0.51	0.5
	word&accent	0.49	0.472	0.471
	word&pause	0.463	0.446	0.433
	All	0.51	0.5	0.5
SVM	Words	0.49	0.48	0.49
	word&accent	0.51	0.49	0.49
	word&pause	0.478	0.453	0.423
	All	0.542	0.52	0.511
MP	Word	0.314	0.432	0.364
	word&accent	0.47	0.46	0.46
	word&pause	0.5	0.473	0.474
	All	0.524	0.53	0.521
LR	Words	0.51	0.5	0.493
	word&accent	0.472	0.46	0.46
	word&pause	0.47	0.46	0.453
	All	0.493	0.49	0.486

in which the *tpngs* pronoun whose function must be predicted, is the third speech token. A speech token can be a word, an hesitation or a pause. The context of each pronoun consists therefore of two speech tokens preceding the pronoun and three speech tokens following it. The data set types used in training are the following: 1) only word tokens, 2) word tokens with stress information, 3) speech tokens consisting of words, hesitation, silent or filled pauses, 4) all features, that is word tokens with eventual stress, hesitation or pause tokens. In **Table 4** a line from each data set is shown in order to illustrate the six-grams language models on which the pronouns were trained. In the data, stress is indicated by the string *ACC* preceding the stressed vowel e.g., *dACCet* is the marked pronoun *det*, while silent pauses are indicated as *SILPAUSE* and filled pauses are joined in the class *FILPAUSE*.

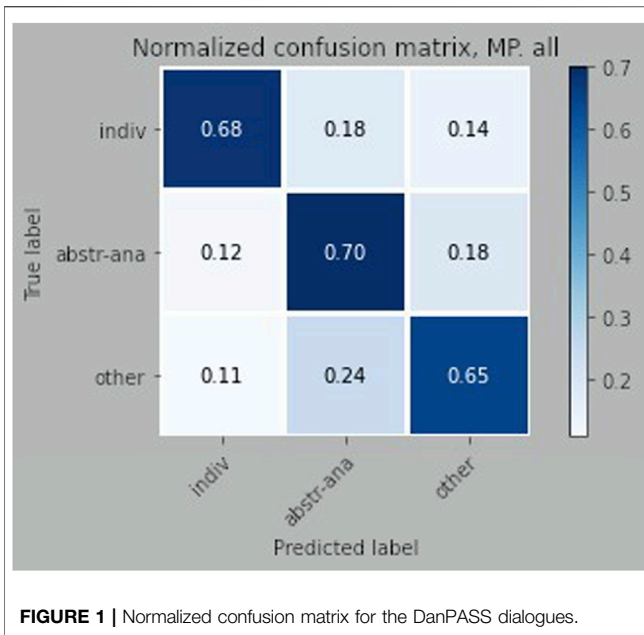
Differing from the experiments in Navarretta (2007), Navarretta (2010), stress information on the pronouns is kept in all data sets, since the usefulness of stress information on the pronouns has been demonstrated in other studies. This change allows us to compare exclusively the impact of the contextual information of the pronouns on classification.

In all the experiments, the baselines are a majority classifier, which chooses the most frequent class and a random classifier,

which assigns a class randomly taking account of the frequency of the classes. The most frequent class in the map task dialogues is *individual*, while it is *other* in the first encounters.

The results of classification on the DanPASS data sets is in **Table 5**, and the results of classification on the NOMCO data sets are in **Table 6**. Precision (P), Recall (R) and weighted F1-score (F1) for each classifier and each data set is shown in the table.

All algorithms perform significantly better than both the majority and random classifiers, but their performance varies from data set to data set. The best results in both map task and first encounters dialogues were obtained by the Multilayer Perceptron trained on the data set including all prosodic features. On the DanPASS data the F1-score is 0.684. The second best performing classifier, on this data, is the Multinomial Naive Bayes Classifier which also performs well on the other data sets. The F1-score by the Multilayer Perceptron trained on the data with all features improves by 0.37 the results of the random classifier while the improvement with respect to the majority baseline is of 0.464. It must be noted that introducing information on pauses alone does not improve classification with respect to data where stress information or only words are used in many cases because the tokens consisting of words and stressed words in the context of pronominal anaphors are more discriminative than the two tokens *filled* or *silent* pause.



However, when both stress information and pauses are used, the best results are obtained. The normalized confusion matrix returned by the Multilayer Perceptron trained on the DanPASS data with stress and pause information is in **Figure 1**. The confusion matrix shows that the class that is predicted correctly more often is that of abstract reference. This is interesting since abstract anaphors are not the most frequent category in the data, and they are difficult to identify without looking at the contextual content Webber (1988), Fraurud (1992), Eckert and Strube (2001). For this reason, abstract anaphors are often excluded from coreference resolution systems. Moreover, the classifier also succeeded in distinguishing abstract and individual anaphors from other types of pronominal uses of *tpngs* pronouns. The averaged results of the ten-fold cross validation gives an F1-score of 0.641, showing that the performance falls only slightly.

The most frequently misclassified functions of pronouns are *other* uses that are classified as *abstract*. A nearer analysis of the wrong classified cases shows that especially cataphoric, and in less degree, abandoned uses of pronouns are misclassified as abstract, if the context included in the language model is not sufficient to discriminate the different uses. Similarly the misclassification between abstract and individual anaphors is mostly due to ambiguous cases in which only the larger context of the dialogue can, in most cases, lead to the correct interpretation of the pronouns. One example is the utterance *det var godt* (it was fine/good) that can both refer to an individual object, in same case independently from the gender of the antecedent, and an abstract entity depending on the context. In some cases, utterances of this type are ambiguous also for humans, and some utterances can have both readings. Only the preferred interpretation was annotated in the data.

Also in the case of the NOMCO data, all classifiers perform better than the two baselines. The best results are obtained by the

Multilayer Perceptron with an F1 score of 0.521. This result improves with 0.35 the majority baseline, and of 0.19 the random baseline. One of the reasons for the lower results obtained on the spontaneous dialogues is that the individual anaphors that are the easiest class to identify are less frequent than the abstract anaphors, while the most frequently occurring pronominal type is the most ambiguous one, the unstressed *det*, which can have all the three functions with nearly the same frequency. The second best performed classifiers is also here the Multinomial Naive Bayes Classifier, which performs best when trained on word token six-grams and on all types of tokens. The averaged ten-fold cross validation gave an higher F1 score = 0.534.

The normalized confusion matrix returned by the Multilayer Perceptron with stress and pause information is in **Figure 2**. The confusion matrix shows that the class that is in most cases classified correctly is *other*. Abstract anaphors and individual anaphors are identified correctly with the same frequency, over 50% of the cases. The classes that are more often confused are *individual* and *other*. Looking at the misinterpreted occurrences, we found that it is especially cataphoric and abandoned uses of the unstressed pronoun *det*, which are confused with its individual uses and vice versa. This happens especially when the pronominal context contains a sequel of so called clausal adverbials, which in Danish precede or follow the finite verb in a clause depending on whether the clause is main or subordinate. The presence of these adverbials results in identical contexts for many types of pronominal uses. Examples of these adverbs in Danish are *da* (surely), *jo* (certainly), *vel* (presumably), *aldrig* (never), *ikke* (not). They can be combined in different ways and are frequent in especially spoken language. The presence of clausal adverbials in the six-grams language models is also one of the reasons behind some of the errors confounding abstract and individual anaphors. Moreover, cases in which only the larger context of the dialogue can indicate the correct interpretation of the pronoun, were found also in these data.

The last classification experiment aimed to determine the type of referent of the abstract anaphors in the DanPASS data. This information has not been annotated yet in the NOMCO data. The

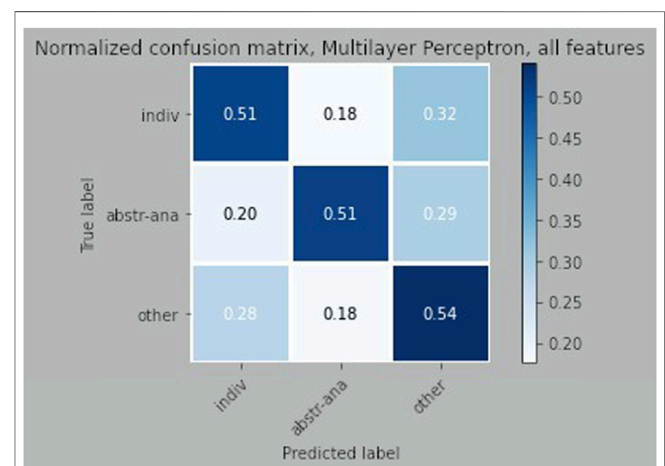


TABLE 7 | Results of classifiers predicting the referent type of abstract anaphors in the DanPASS dialogues.

Classifier	Dataset	P	R	F1
Majority		0.175	0.419	0.247
Random		0.274	0.209	0.23
KNC	Words	0.57	0.581	0.564
	words&accent	0.597	0.605	0.596
	words&pause	0.638	0.605	0.6
	All	0.639	0.628	0.617
MNB	Word	0.637	0.628	0.616
	word&accent	0.672	0.651	0.644
	word&pause	0.672	0.674	0.668
	All	0.71	0.698	0.698
SVM	Words	0.602	0.604	0.6
	word&accent	0.693	0.674	0.671
	word&pause	0.678	0.674	0.67
	All	0.707	0.698	0.67
MP	Word	0.68	0.7	0.69
	word&accent	0.638	0.651	0.644
	word&pause	0.622	0.651	0.631
	All	0.734	0.744	0.737
LR	Words	0.6	0.674	0.665
	word&accent	0.692	0.698	0.69
	word&pause	0.684	0.698	0.69
	All	0.733	0.744	0.736

data consists of 211 six-grams as in the preceding experiments, but only abstract pronouns have been held and information about their referent type is added as the class to be predicted. The ambiguous class “event-fact” is treated as an eventuality, thus the four classes to be predicted are *eventuality*, *fact-like*, *proposition* and *speech-act*. The best data set from the preceding experiment, that is the language model which contains all prosodic information was used. In **Table 7**, we present the results of the two baselines and the two best performing algorithms, which on this task and data are the K Neighbors Classifier (KNC) and the Multilayer Perceptron run with the following hyper-parameters: The *sgd* solver with the *relu* activation, two layers of size 6, and 1, 6,000 iterations, adaptive learning rate and alpha = 0.0001. Also in this table the results are presented in the form of weighted Precision, Recall and F1 score.

The best results were obtained again by the Multilayer Perceptron. The F1 score was 0.737, which is an improvement of 0.49 respect to the best performing baseline, in this case the majority classifier, and of 0.507 respect to the random classifier. Given that it is often difficult to determine the referent type also for humans, these results are very positive. In this case, most errors were confusing eventuality and fact readings of the pronoun. The F1 score from ten-fold cross validation is 0.71.

We also repeated the classification experiment, which gave the best results in Navarretta (2007) and that it is discussed in Navarretta (2010) in order to test whether the use of filled pauses decreases or increases the classification results. The original experiment was run in the WEKA SMO classifier (a support vector machine) with ten-fold cross validation and the aim was to classify all nine functions of the *tpngs* pronouns from the DanPASS dialogues. In the experiment, the effect of pauses was only measured together with stress information added to

words, and the F1 score reported in Navarretta (2007) is 0.518. We run the scikit-learn SVM classifier with a linear kernel on the extended data set and obtained an F1-measure of 0.554. Running ten-fold cross-validation gave an F1-score of 0.549. These results show that information about filled pauses improves classification, even if there are few of them.

6 DISCUSSION

The analysis of the dialogue annotations show that even if personal and demonstrative *tpngs* pronouns are used frequently in the dialogues with both an individual and abstract anaphoric function, there is a preference for the marked pronouns to have an abstract antecedent and for the unmarked personal pronoun *det* to have individual antecedents. The preference is stronger in the first encounters than in the map task dialogues. Thus, the observation made by Navarretta (2004), Navarretta (2007) that in Danish the personal pronoun *det* is the preferred anaphor in both individual and abstract reference only holds in written language, where it is not possible to distinguish between unmarked and marked occurrences of the pronoun.

The analysis of the DanPASS dialogues also shows that both silent and filled pauses precede more frequently abstract anaphors than individual anaphors. This indicates that speech pauses can signal that the speaker is going to utter a difficult concept, since according to all theories of reference *tpngs* pronouns with abstract antecedents are less salient/accessible to the addressee than pronouns with individual antecedents, and they are also more difficult to express for the speaker. This work is therefore in line with perception studies that found that speakers use more frequently pauses when they have to utter or define abstract concepts than concrete ones Rochester (1973), Reynolds and Paivio (1968).

We also found that unmarked pronouns are the most common anaphors with a proposition referent even if propositions are the most abstract types of entities according to e.g., (Asher, 1993). On the other hand, in the DanPASS dialogues, marked anaphors are the most frequently occurring pronouns with referents of the eventuality type, which according to researchers investigating reference in English e.g., Webber (1988), Gundel et al. (1993), Asher (1993) are the less abstract type of abstract entity, and they are therefore often referred to by personal pronouns. The analysis of pauses preceding the abstract anaphors in our data also shows that personal pronouns with a fact or proposition referent are more often preceded by pauses than when they refer to eventualities. We propose that the presence of a pause, in these cases, might mark that the referent of the pronoun is not the most expected one. In this cases, pauses have a similar function as that observed by Gargiulo et al. (2019) on Italian and Swedish overt individual pronouns with subject or object antecedents. However, since the data are of limited size and we do not have other corpora to compare our data with, this supposition should be tested in more dialogues. Moreover, we did not notice a similar use of pauses when marked pronouns referred to eventualities. In general, with respect to reference, it is only possible to study preferred uses since other factors than salience

can move a speaker to use one form of reference instead of others, some of these being variation and personal preferences. In the future, we will also analyze the use of pauses in the first encounters.

The machine learning experiments aimed to determine to what extent individual, abstract anaphors and other uses of Danish *tpngs* pronouns can be predicted automatically training classifiers on language models of speech tokens were strongly inspired by the experiments performed by Navarretta (2007) on the same data. For example, the best performing language models in those experiments, six-grams, were used in the present work.

The experiments were also repeated on new annotated data, part of the spontaneous Danish NOMCO first encounters corpus. The results of our experiments show that a semi-automatically tuned Multilayer Perceptron can identify the correct function of *tpngs* pronouns in more than two thirds of their occurrences, with a weighted F1-score of 0.67 on the DanPASS data, and on 0.52 of the cases on the first encounters. Running ten-fold cross validation on the same data sets gave similar results. These are both significant improvements of the F1 score achieved by the random and majority classifiers. Interestingly, the class which is identified most correctly in the map task dialogues is that of abstract anaphors, even if they are not the most frequently occurring type in these data.

In the first encounters, the class that is identified correctly in more cases is the *other* class, which is the most frequent one. However, the confusion matrix from this experiment indicates that also abstract and individual uses of *tpngs* pronouns are correctly identified in over 50% of the cases. The analysis of erroneous classified entities shows that in the map task dialogues the classes most often confused are *other* and abstract anaphors. In particular cataphoric and abandoned pronouns are often confused with abstract anaphors and vice versa when the context included in the used language model is not large enough. In the first encounters, abstract anaphors and cataphoric or abandoned pronouns are the classes that are most often confused by the classifiers. Finally, insufficient contextual size was the cause of many errors classifying abstract anaphors as individual ones and vice versa.

The obtained results indicate that using information on speech pauses and stress on words can contribute significantly to the task of distinguishing individual, abstract pronominal anaphors, and other pronominal functions automatically. The automatic classification could replace or support rule-based discrimination of individual and abstract anaphors in anaphora resolution systems.

The fact that introducing information on pauses alone does not improve classification with respect to data where stress information or only words are used is due to the fact that tokens consisting of words and stressed words surrounding pronominal anaphors are more discriminative than the two tokens *filled* or *silent* pause. However, when pauses are added to words with stress information the classification results improve for most of the classifiers. It most also been noticed that running ten-fold cross classification with the Multilayer Perceptron gives a fall of approx. 0.04 in performance with respect to when we run the experiments training the data on 80% of the data, but they are

still good indicating that the results are quite reliable on these data at least.

The best measure we have in order to compare our classification results with the current state of art is given by the results reported by Marasovic et al. (2017) for the resolution of English *tpngs* pronouns in the WSJ part of the ARRAU corpus. The results were obtained by LSTM trained on the many morphological and syntactic features in the corpus annotations. The precision of the machine learning algorithm is reported to be 29.01. They also report that the algorithm proposed the correct antecedent as the fourth ranked candidate in the antecedent candidate list in 63.55 of the cases. In these candidate list were often both nominal and verbal phrases, as well as clauses. Our results are not directly comparable, since Marasovic et al. (2017) not only identify, but also resolve English abstract and individual *tpngs* pronouns in approx. one third of their occurrences. However, the difference in performance on the resolution of *tpngs* pronouns compared to that obtained by coreference systems in general, shows clearly how difficult the present task is. Moreover, our results are interesting for different reasons. First of all, the Danish unmarked personal pronoun is used much more often than in English to refer to abstract entities when we compare our data with what has been reported for English e.g., Webber (1988), Gundel et al. (1993), Poesio and Artstein (2008). Therefore, abstract and individual anaphors in Danish are harder to discriminate on the basis of the pronominal type than in English. Secondly, our data are dialogues, which are notoriously more difficult to process than texts, while Marasovic et al. (2017) address newspapers. Finally, we have not relied on morphological and syntactic features as anaphor resolution systems do by choice, since our aim has been to investigate the importance of pauses and stress for distinguish between different functions of the Danish *tpngs* pronouns.

The six-grams language models were also used in order to identify the referent type of the abstract anaphors in the DanPASS data. This task has not been attempted on other languages, and it is also difficult for humans as discussed in e.g., Webber (1991), Fraurud (1992). We achieved good results, F1 score = 0.737, but we only run the classifiers on abstract anaphors. However, knowing the type of referent is useful to identify the correct antecedent of abstract anaphors, and this information is not present in the English corpora including abstract pronominal anaphora annotations.

7 CONCLUSIONS AND FUTURE STUDIES

In the paper, we presented a study of individual and abstract anaphoric reference and the role of speech pauses that precede them in two annotated types of Danish dyadic dialogues: map task dialogues and first encounters. The study has given new insight into the phenomenon of pronominal individual and abstract reference in Danish spoken language, pointing out the role of speech pauses in the task of determining whether *tpng* pronouns are used anaphorically or not and, in the case they are

anaphors, whether they refer to an individual or an abstract entity.

This work revises preceding studies that concluded that the most common abstract pronoun in Danish is the unmarked pronoun *det* (Navarretta, 2010), since this is only valid for Danish texts. In both types of dialogues, in fact, there is a statistically significant preference for referring to individual entities with the unmarked personal pronoun *det*, while the stressed *d'et* and the demonstrative pronouns *det her* and *det der* have more often abstract referents as it is also the case for the corresponding pronouns in English. Secondly, we found that silent and filled pauses precede much more frequently abstract pronominal anaphors than individual pronominal anaphors, confirming preceding studies that showed that people produce more hesitations and silent pauses when they are going to utter an abstract concept than a concrete one (Rochester, 1973). Thirdly, on the basis of the analysis of speech pauses preceding abstract pronominal anaphors with different types of referents, we propose tentatively that speech pauses signal that the referent of an unmarked anaphor is of a type less expected given the pronominal type.

Finally, our machine learning experiments confirm the fact that information about occurrences of speech pauses and stress information on words can classify individual anaphors, abstract anaphors or other functions of Danish *tpsng* pronouns in more than two thirds of their occurrences in the map task dialogues and in more than half cases in the first encounters. Adding information on filled pauses also helps classifying all uses of

tpsng pronouns improving the experiments in (Navarretta, 2007).

Since the proposals in this paper are based on the analysis of one types of dialogue, the use of pronominal anaphors in spoken data and the function of pauses with respect to individual and abstract reference should be investigated in more spontaneous dialogues and in different domains. In the future, it should also be tested whether silence pauses can signal the occurrences of abstract anaphors or that the pronoun they precede has a referent of a more abstract type than that pre-announced by the pronominal type in other languages.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The DanPASS dialogues are publicly available, the DAD data and the NOMCO annotations can be obtained contacting the author. Requests to access these datasets should be directed to DanPASS, <https://danpass.hum.ku.dk/>, DAD and NOMCO annotations to costanza@hum.ku.dk.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

- Allan, R., Holmes, P., and Lundskaer-Nielsen, T. (1995). *Danish - A Comprehensive Grammar*. London: Routledge.
- Allwood, J. (1988). "The Structure of Dialog," in *Structure of Multimodal Dialog II*. Editors M. M. Taylor, F. Neél, and D. G. Bouwhuis (Amsterdam: John Benjamins), 3–24.
- Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., et al. (1991). The Hrc Map Task Corpus. *Lang. Speech* 34, 351–366.
- Ariel, M. (2001). "Accessibility Theory: An Overview," in *Text Representation, Human Cognitive Processing Series*. Editors T. Sanders, J. Schlieroord, and W. Spooren (John Benjamins), 29–87.
- Ariel, M. (1988). Referring and Accessibility. *J. Linguistics* 24, 65–87.
- Asher, N. (1993). "Reference to Abstract Objects in Discourse," in *Studies in Linguistics and Philosophy* (Dordrecht, Netherlands: Kluwer Academic Publishers), Vol. 50.
- Barhom, S., Schwartz, V., Eirew, A., Bugert, M., Reimers, N., and Dagan, I. (2019). "Revisiting Joint Modeling of Cross-Document Entity and Event Coreference Resolution," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Florence, Italy: Association for Computational Linguistics), 4179–4189.
- Boersma, P., and Weenink, D. (2009). Praat: Doing Phonetics by Computer. Available at: <http://www.praat.org/> (Retrieved May 1, 2009).
- Byron, D. K. (2002). "Resolving Pronominal Reference to Abstract Entities," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02) (Philadelphia, PA: Association of Computational Linguistics), 80–87.
- Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistics. *Comput. Linguist.* 22, 249–254.
- Chafe, W. (1987). "Cognitive Constraint on Information Flow," in *Coherence and Grounding in Discourse*. Editor R. R. Tomlin (Amsterdam: John Benjamins), 20–51.
- Clark, H. H., and Fox-Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition* 84, 73–111. doi:10.1016/s0010-0277(02)00017-3
- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* 20, 37–46. doi:10.1177/001316446002000104
- Duncan, S., and Fiske, D. (1977). *Face-to-Face Interaction*. Hillsdale, NJ: Erlbaum.
- Eckert, M., and Strube, M. (2001). Dialogue Acts, Synchronising Units and Anaphora Resolution. *J. Semantics* 17, 51–89. doi:10.1093/jos/17.1.51
- Esposito, A., Duncan, S., and Quek, F. (2002). "Holds as Gestural Correlated to Empty and Filled Pauses," in Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002) (Denver, Colorado: ISCA), Vol. 1, 541–544.
- Esposito, A., and Esposito, A. M. (2011). "On Speech and Gesture Synchrony," in *Communication and Enactment - The Processing Issues, LNCS*. Editors A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, and A. Nijholt (Stockholm, Sweden: ACL), Vol. 6800, 252–272.
- Fraurud, K. (1992). *Processing Noun Phrases in Natural Discourse*. Stockholm, Sweden: Department of Linguistics - Stockholm University.
- Gargiulo, C., Tronbier, M., and Bernardini, P. (2019). The Role of Prosody in Overt Pronoun Resolution in a Null Subject Language and in a Non-Null Subject Language: A Production Study. *Glossa: A J. Gen. Linguist.*, 135–156. doi:10.5334/gjgl.973
- Givón, T. (1979). *On Understanding Grammar*. New York, NY: Academic Press.
- Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in Spontaneous Speech*. London: Academic Press.
- Grönning, N. (2009). A Danish Phonetically Annotated Spontaneous Speech Corpus (DanPASS). *Speech Commun.* 51, 594–603. doi:10.1016/j.specom.2008.11.002
- Research Challenges in Speech Technology: A Special Issue in Honour of Rolf Carlson and Björn Granström
- Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive Status and the Form of Referring Expressions in Discourse. *Language* 69, 274–307.
- Hirschberg, J., and Nakatani, C. (1998). "Acoustic Indicators of Topic Segmentation," in Proceedings of ICSLP-98 (ISCA: Sidney).

- Kameyama, M. (1998). "Intrasentential Centering: A Case Study," in *Centering Theory in Discourse*. Editors M. Walker, A. Joshi, and E. Prince (Oxford, UK: Oxford University Press), 89–112.
- Kolhatkar, V., Roussel, A., Dipper, S., and Zinsmeister, H. (2018). Survey: Anaphora with Non-Nominal Antecedents in Computational Linguistics: A Survey. *Comput. Linguist.* 44, 547–612. doi:10.1162/colia00327
- Krauss, R., Chen, Y., and Gottesman, R. F. (2000). "Lexical Gestures and Lexical Access: A Process Model," in *Language and Gesture*. Editor D. McNeill (Amsterdam, Netherlands: John Benjamins), 261–283.
- Lyons, J. (1977). *Semantics*. Cambridge University Press, Vols. I–II.
- Maclay, H., and Osgood, C. E. (1959). Hesitation Phenomena in Spontaneous English Speech. *Word* 15, 19–44.
- Marasovic, A., Born, L., Opitz, J., and Frank, A. (2017). "A Mention-Ranking Model for Abstract Anaphora Resolution," in Proceedings of EMNLP 2017 (Copenhagen, Denmark: Association of Computational Linguistics), 221–232.
- Müller, C. (2007). "Resolving it, This and that in Unrestricted Multi-Party Dialog," in Proceedings of ACL-2007 (Prague: ACL), 816–823.
- Navarretta, C. (2007). "A Contrastive Analysis of Abstract Anaphora in Danish, English and Italian," in Proceedings of DAARC 2007 Editors A. Branco, T. McEnery, R. Mitkov, and F. Silva (Lagos, Portugal: Centro de Linguística da Universidade do Porto), 103–109.
- Navarretta, C., Ahlsén, E., Allwood, J., Jokinen, K., and Paggio, P. (2012). "Feedback in Nordic First-Encounters: A Comparative Study," in Proceedings of LREC 2012 (Istanbul, Turkey), 2494–2499.
- Navarretta, C., and Olsen, S. (2008). "Annotating Abstract Pronominal Anaphora in the DAD Project," in Proceedings of LREC-2008 (Marrakesh, Morocco: ELRA), 2046–2052.
- Navarretta, C. (2004). "Resolving Individual and Abstract Anaphora in Texts and Dialogues," in COLING-2004: Proceedings of the 20th International Conference of Computational Linguistics (Geneva, Switzerland), 233–239.
- Navarretta, C. (2010). Stress, Pauses, Pronominal Types and Pronominal Functions in Danish Spoken Data. *Copenhagen Stud. Lang.*, 45–60.
- Navarretta, C. (2002). The Use and Resolution of Intersentential Pronominal Anaphora in Danish Discourse. PhD thesis. Copenhagen, Denmark: University of Copenhagen.
- Orasan, C. (2003). "PALinkA: A Highly Customizable Tool for Discourse Annotation," in Proceedings of the 4th SIGdial Workshop (Sapporo, Japan: ACL), 39–43.
- Paggio, P., and Navarretta, C. (2017). The Danish NOMCO Corpus of Multimodal Interaction in First Acquaintance Conversations. *Lang. Resour. Eval.* 51, 463–494. doi:10.1007/s10579-016-9371-6
- Poesio, M., and Artstein, R. (2008). "Anaphoric Annotation in the ARRAU Corpus," in Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08) (Marrakech, Morocco: European Language Resources Association (ELRA), 1170–1174.
- Poesio, M., Grishina, Y., Kolhatkar, V., Moosavi, N., Roesiger, I., Roussel, A., et al. (2018). "Anaphora Resolution with the ARRAU Corpus," in Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference (New Orleans, Louisiana: Association for Computational Linguistics), 11–22. doi:10.18653/v1/W18-0702
- Poesio, M. (2004). "The Mate/gnome Proposals for Anaphoric Annotation, Revisited," in Proceedings of the 5th SIGdial Workshop Editors M. Strube and C. Sidner (Cambridge, Massachusetts, United States: Association for Computational Linguistics), 154–162.
- Prince, E. F. (1992). "The ZPG Letter: Subjects, Definiteness, and Information-Status," in *Discourse Description. Diverse Linguistic Analyses of a Fund-Raising Text*. Editors W. Mann and S. A. Thompson (Amsterdam, Netherlands: John Benjamins), 295–325.
- Prince, E. F. (1981). "Toward a Taxonomy of Given-New Information," in *Radical Pragmatics*. Editor P. Cole (New York, NY: Academic Press), 223–255.
- Reynolds, A., and Paivio, A. (1968). Cognitive and Emotional Determinants of Speech. *Can. J. Psychol.* 22, 164–175.
- Rochester, S. R. (1973). The Significance of Pauses in Spontaneous Speech. *J. Psycholinguistic Res.* 2, 51–81.
- Roesiger, I., and Riester, A. (2015). "Using Prosodic Annotations to Improve Coreference Resolution of Spoken Text," in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (Beijing, China: Association for Computational Linguistics), 83–88. doi:10.3115/v1/P15-2014
- Roesiger, I., Stehwi, S., Riester, A., and Vu, N. T. (2017). "Improving Coreference Resolution with Automatically Predicted Prosodic Information," in Proceedings of the Workshop on Speech-Centric Natural Language Processing (Copenhagen, Denmark: Association for Computational Linguistics), 78–83. doi:10.18653/v1/W17-4610
- Shriberg, E. (1994). Preliminaries to a Theory of Speech Disfluencies. PhD thesis. Berkeley: University of California.
- Strube, M., and Müller, C. (2003). "A Machine Learning Approach to Pronoun Resolution in Spoken Dialogue," in Proceedings of the ACL'03, 168–175.
- Uryupina, O., Artstein, R., Bristot, A., Cavicchio, F., Delogu, F., Rodriguez, K. J., et al. (2020). Annotating a Broad Range of Anaphoric Phenomena, in a Variety of Genres: The Arrau Corpus. *Nat. Lang. Eng.* 26, 95–128. doi:10.1017/S1351324919000056
- Vendler, Z. (1963). The Grammar of Goodness. *Philos. Rev.*, 446–465.
- Webber, B. (1988). Discourse Deixis and Discourse Processing. Tech Report. University of Pennsylvania.
- Webber, B. L. (1991). Structure and Ostension in the Interpretation of Discourse Deixis. *Nat. Lang. Cogn. Process.* 6, 107–135.
- Yang, B., and Mitchell, T. M. (2016). "Joint Extraction of Events and Entities within a Document Context," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (San Diego, California: Association for Computational Linguistics), 289–299. doi:10.18653/v1/N16-1033

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.

Copyright © 2021 Navarretta. 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.



Experiment in a Box (XB): An Interactive Technology Framework for Sustainable Health Practices

m. c. schraefel^{1*}, George Catalin Muresan¹ and Eric Hekler²

¹WellthLab, Electronics and Computer Science, University of Southampton, Southampton, United Kingdom, ²Herbert Wertheim School of Public Health and Human Longevity Sciences, University of California, San Diego, CA, United States

OPEN ACCESS

Edited by:

Gennaro Cordasco,
University of Campania Luigi Vanvitelli,
Italy

Reviewed by:

Miriam Sturdee,
Lancaster University, United Kingdom
Mike Fraser,
University of Bath, United Kingdom

*Correspondence:

m. c. schraefel
mc+w@ecs.soton.ac.uk

Specialty section:

This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

Received: 31 January 2021

Accepted: 06 May 2021

Published: 26 May 2021

Citation:

schraefel mc, Muresan GC and
Hekler E (2021) Experiment in a Box
(XB): An Interactive Technology
Framework for Sustainable
Health Practices.
Front. Comput. Sci. 3:661890.
doi: 10.3389/fcomp.2021.661890

This paper presents the Experiment in a Box (XB) framework to support interactive technology design for building health skills. The XB provides a suite of experiments—time-limited, loosely structured evaluations of *health heuristics* for a user-as-experimenter to select from and then test in order to determine that heuristic's efficacy, and to explore how it might be incorporated into the person's life and when necessary, to support their health and wellbeing. The approach leverages self-determination theory to support user autonomy and competence to build actionable, personal health *knowledge skills and practice* (KSP). In the three studies of XB presented, we show that with even the short engagement of an XB experiment, participants develop health practices from the interventions that are still in use long after the intervention is finished. To situate the XB approach relative to other work around health practices in HCI in particular, we contribute two design continua for this design space: *insourcing to outsourcing* and *habits to heuristics*. From this analysis, we demonstrate that XB is situated in a largely under-explored area for interactive health interventions: the insourcing and heuristic oriented area of the design space. Overall, the work offers a new scaffolding, the XB Framework, to instantiate time-limited interactive technology interventions to support building KSP that can thrive in that person, significantly both post-interventions, and independent of that technology.

Keywords: embodied interaction, insourcing, outsourcing, continua, embodied, knowledge skills and practice, ksp, heuristics

INTRODUCTION

In this paper, we present the Experiment in a Box framework, or XB for short, for designing interactive tools to support a person building *knowledge skills and practice* (KSP) to maintain their own health and wellbeing. By knowledge, we mean awareness, understanding, and wisdom related to the one's body as it constantly adapts across time and in context toward enabling a person to move toward desired adaptive states (e.g., being fit, reducing stress, maintaining a healthy weight, etc). By skills, we are referring to one's capacity to regulate and “tune” oneself in and to context toward desired adaptive states (schraefel and Hekler, 2020). By practice, we are referring to two senses of the term: repetition and commitment. Practice in the first sense is the number and frequency of the repetitions of a skill necessary to become adept at the knowledge and skills necessary in this case to build and maintain health. That repetition, in the second sense, as a formal *practice* frames this repetition as an ongoing part of one's life, of actively, continually and deliberately engaging in

developing and refining desired knowledge and skills toward desired states. The XB, as we detail in Related Work below, addresses a gap in the research around how to design interactive technology to help people build these resilient (Feldman, 2020) adaptable health KSP that can be actioned across health-challenging contexts, without requiring that technology to be perpetually necessary. In other words, we want to design technology to support building health KSP so people can get off that technology, and thrive with those KSP skills. Drawing on recent work that suggests framing health technology within a continuum from *outsourcing* to *insourcing* health skills (schraefel et al., 2020), we show that most of the current research and commercial interactive health technology supports the *outsourcing* end of health tech design—that is, giving the problem to a third party to manage, including especially around health, to automated interactive technology devices and services, from step counters or follow along workouts and their associated data capture and visualisations, to coordinated product and service delivery.

Further, within the increasingly pervasive discourse of *behavior change/habit building* frame, many of the current technologies not only seek to outsource habit-building but to then also seek to cultivate a person's reliance not just on the technology but also context to drive these habits-as-healthy behaviour-change. For example, it is increasingly common to outsource one's exercise to a follow-along program, or in some cases to programs that also monitor various biometrics, where tools suggest strategies based on that data for making progress ("TrainingPeaks | New Year. New Focus." n. d.; "The Sufferfest: Complete Training App for Cyclists and Triathletes" n. d.). While there is great potential for such outsourced, habit-formation-oriented technologies, to date, the hope has not been observed in reality. (Riley et al., 2011; West et al., 2012; Lunde et al., 2018; McKay et al., 2018). A key reason for this gap is that an outsourced, habit-formation orientation to behavior change asks for a very heavy lift for both the technology and for context to drive a person into desired adaptive states. This is difficult based on the inherent complexity of human behavior and health, which has been summarized elsewhere in terms of recognizing three factors that impact behavior and health: context, timing, and individual differences (Hekler et al., 2019; Chevance et al., 2020; Hekler et al., 2020; ;). These dynamics require what we frame as *constant adaptivity in context* (schraefel 2020). These complexities establish the need for outsourced, habit-formation oriented technologies to be highly adaptive to context and changing circumstances—which is somewhat the antithesis of habits as context specific, context dependent. Most current systems are not broad enough or expert enough to be sufficiently adaptive, which, we contend, is the primary reason behind the gap between the hope and reality. For instance, most tools cannot be responsive and supportive to a simple question like what to do when it's too late to go for the planned run; what are options when there are no stairs to climb? These current outsourced systems are also high cost in terms of their subscriptions, the associated special gear (Ji et al., 2014), and the need for complex, often data invasive "personalization" algorithms (Hekler et al., 2018; Hekler et al., 2020). Thus, while

there is possibility and promise, considerable work is needed before this form of responsive health practices could become a reality.

An alternative to outsourced health practices/health behavior change is emerging under the auspices of personal science (Wolf and De Groot, 2021) or "self-study" (Nebeker et al., 2020). This approach emphasizes what we call *insourcing*, meaning the cultivation of KSP within the person, thus establishing (and supporting) the person as the driver and lead for achieving desired states, not the technology. The dominant way in which personal science/self-study are being supported *via* technology, particularly in the human-computer interaction (HCI) literature, is through extensions of randomized control trial logic into N-of-1 cross-over trials, which involve randomizing, within person and across time, the delivery or not of an intervention relative to some meaningful comparator (Karkar et al., 2017). These approaches focus on gathering quantifiable data to refine an *outsourced* prescription for health activities, ranging from workout plans (Agapie et al., 2016) to sleep strategies (Daskalova et al., 2016). Implicit in many of these approaches still remains an emphasis on either a single action behavior change, such as avoiding certain foods that appear linked with undesired symptoms, or cultivating a prescribed habit and routines, such as specific sleep hygiene techniques.

Our goal with the XB framework is to extend personal science and insourcing by shifting emphasis from behavior change/habit formation to, instead, cultivation of KSP and what we frame as heuristics. By heuristics, we mean a set of general principles, grounded in prior knowledge from both science and lived experience, that can guide effective adaption of the person, across contexts, toward desired adaptive states. The XB framework seeks to cultivate personal health heuristics by blending the broad categories of *outsourced* health and wellbeing plans and programs (prior knowledge based on science) as initial generic heuristics with the *insourcing* of personal evaluation found in N-of-1 studies in order for a person to test and customize those heuristics for themselves, to work across contexts.

The focus of the XB framework is, therefore, to help build the necessary health KSP so that a person 1) can answer the question "how do you feel" as a meaningful self-diagnostic and 2) from this self-diagnostic, apply associated health and wellbeing KSP toward selecting appropriate actions in context to feel better. As pointed to above, we frame this ability to select and adapt appropriate KSP to build and maintain health and wellbeing across contexts as *tuning* (schraefel and Hekler, 2020). The term draws on the analogy of tuning as an instrument or machine: aligning the components of a mechanism so that they resonate together, reinforcing each other toward better performance. The strings of a guitar, for example, are tuned to particular pitches, harmonically, relative to each other enabling them to produce music within the frames of those harmonics; that instrument itself may be tuned to a standard frequency so that it resonates harmonically with other instruments as well. In a *tuned* engine, all components are likewise synchronized, thus producing more power, more efficiency.

Our goal in presenting the XB framework here is two-fold. First, we seek to expand the design space of health technologies to suggest the possibility for solutions that work across both the

spectrums of outsourcing to insourcing, and across the dimensions of cultivating behavior change/habit formation to cultivating personal health heuristics *via* building and testing associated health KSP. Second, with the XB framework specifically, our goal is to help designers and individuals develop interventions that will support dynamic self-tuning for one's healthful resilience, across contexts. In the following sections, we present a more detailed discussion of the related work motivating the XB approach. We then present three exploratory studies developed to help triangulate on the set of features that are the fundamentals of the XB framework.

These three studies were conducted explicitly to explore the XB model, framed *via* the ORBIT model (Czajkowski et al., 2015) notion of “proof-of-concept” studies. In particular, we follow the ORBIT approach of *Defining, Refining, and Testing* for a meaningful signal. Note that, as discussed in ORBIT, this is an appropriate methodology for this stage of the research process present here. From the studies reported below, we can say convincingly that we appear to be getting a clear signal across these three studies. In accordance with the ORBIT model, this finding can be used to then justify more rigorous evaluation relative to a meaningful compactor. Our studies were explicitly conducted iteratively and consecutively to enable exploration and refinement of the XB approach. The fact that similar conclusions were being drawn even across disparate samples, with the first focused on an ad agency, the second on a university setting, and the third with an open call on social media, point toward some degree of consilience and, thus, increased confidence in the value of continuing this line of work toward more formal and rigorous testing, such as randomized controlled trials.

We conclude by summarizing our findings and identifying opportunities for future work that offer some reflections on open questions about XB phases, and also challenges for use in supporting health exploration for group/cultural health practice and infrastructure, beyond the individual. Overall, the XB framework offers a set of novel parameters, such as insourcing health knowledge skills and practice thus opening new areas in the HCI health design space to be explored.

RELATED WORK

Supporting health and wellbeing is a burgeoning research area in HCI, from addressing depression, anxiety, or bipolar health issues to new models of digitally enabled healthcare delivery e.g., (Blandford, 2019; Sanches et al., 2019). In the following sections situating our work relative to this context, we further develop both the *insourcing to outsourcing* design continuum (schraefel et al., 2020) (schraefel et al., 2021) and the *habits to heuristics* continuum outlined above as new vectors for contextualizing the health design space.

Insourcing to Outsourcing Lens of Interactive Health Tech

Outsourcing (Troacă, 2012) refers to the subcontracting out of expertise or production to reduce costs and improve efficiencies

in industry. Tim Ferris's book, *The Four Hour Workweek* (Ferriss 2009) made the concept more personal: one could outsource to a third party tasks from managing one's calendar to one's laundry to free up one's time for more time-valuable processes. Outsourcing also outsources oversight of these processes, which can lead to issues from cost overruns and poorer quality to abuse, from overpriced contracts and substandard services (Brooks, 2020), to suicides at outsourcing factories (Heffernan, 2013) and inmate abuse, motivating the recent shutting down in the United States of outsourced federal prison services (Sabrina and sadie, 2021).

As a complement to outsourcing, we proposed “insourcing” (schraefel et al., 2020) to consider bringing tasks in-house, restoring them to the entity-as-owner. Such insourcing in physical activity would include a person undertaking to learn how to build skills for themselves to achieve and maintain fitness, rather than relying on a trainer to manage programming their fitness experience. Insourcing also has costs: time and money to gain skills, funds for associated equipment, time and resources to practices skills; risks that one's expertise may not be sufficient on its own to support good strategies for any associated goals. On the plus side, insourced knowledge skills and practice can enable one to be more responsive to challenges, for much longer. A recent workshop call (schraefel et al., 2021), focusing on exploring the insourcing end of the insource-outsource continuum makes clear that there is not a value judgment in insourcing or outsourcing. Insourcing \leftrightarrow Outsourcing is, instead, a continuum: we often flow between these points. For example, most of us outsource production, management and distribution of vegetables from field to home to grocery stores—we outsource that expertise, time and risk and accept its limitations in choice and quality. We may also insource some small part of that process from time to time, as amateur farmers, working on urban community gardens, or growing a tomato in a pot over the summer, sufficient to dress the occasional salad.

This continuum of insourcing to outsourcing is one lens we can use to interpret emphasis in health tech both commercially and in interactive tech for health research. As we show in the following section, while there is a spread across the continuum, the preponderance of work clusters around outsourcing. Through the pandemic, such outsourcing has only escalated: sales of Peloton bikes and subscriptions to associated follow-along anytime-workouts have significantly increased since pre-COVID (Griffith, 2020). Subscription-based guided health services from yoga to meditation have also risen (Lerman, 2020). A trait also strongly associated with these approaches is quantitative data tracking, where sensors capture biometric information while associated apps translate these into scores or performance ratings. Step counters like Fitbit hardware and associated apps are a great example of this approach: the user tacitly agrees to follow the advice of the device which offers a simple, understandable directive: “get X steps on this counter within a day”. It is easy for a person to determine success or failure. Over the past 5 years in particular, more sensors have been integrated into single devices, from phones to watches, gathering even more types of biometric data, less obtrusively: one watch can capture heart rate, standing/moving time, sleep

time movement pace, incorporated of course with where and when each beat and step is taken¹. These measures offer a happy blending between computer scientist competence and the medical communities' methodologies. As computer scientists, we are very good at creating ever cheaper, lower power, longer-lasting finer grained sensors. In the medical and sports science communities, we regularly make our cases on health prescription by connecting quantified measures with outcomes: blood pressure, heart rate, age, girth correlate to outcomes from disease prevention approaches [e.g. (Barry et al., 2014; Stevens et al., 2016)] to athlete training programs (Allen et al., 2019).

Beyond the off-the-shelf apps and tools like sports watches and health tracking apps that clearly define what can be counted and correlated, the Quantified Self (QS) communities have demonstrated the value of being able to correlate an amazing variety of measures to *see* more consistent data to both understand and adapt more varied practices to support desired outcomes (Prince, 2014; Barcena et al.), from room temperature to blood glucose, heart rate, weight as part of deliberate self-study. This QS community is heterogeneous. While some QS-oriented people collect more or less all kinds of data (Ajana, 2017), and some have more specific goals, such as monitoring sports performance over years, to help reflect on their practice or tune it (Dulaud et al., 2020; Heyen, 2020), others carry out more deliberate experiments, usually to understand a possible correlation. That is: a formal protocol is developed; it is executed over a specific time; the data is gathered; and then analyzed toward a conclusion (Choe et al., 2015a).

In more formal research in HCI, the outcome of such exploration has often been to support informing an outsourced outcome: a specific health prescription (Consolvo et al., 2008), supporting data-derived insights (Bentley et al., 2013) and/or using self-experimentation [e.g. (Hekler et al., 2013; Agapie et al., 2016; Duan et al., 2013; Daskalova et al., 2016; Pandey et al., 2017; Lee et al., 2017; Kravitz et al., 2020)] to support a decision to adopt a practice. Early days in this research (circa 2008) provided largely prescriptive interventions, with research focused on persuasive technology approaches of how to design the interaction to support translating a person's actions with the intervention into a clearly prescribed habit, such as "walk 10 k steps"; "stand every hour"; "go to the gym" [e.g. Ubifit (Consolvo et al., 2008)]. Over the past decade, inspired by QS, personal informatics and n-of-1 studies, health practice designs in self-study have moved toward focusing increasingly on using one's own data for personal sense-making. Bentley et al.'s (Bentley et al., 2013) Health Mashups attempts to find and highlight possible correlations between quantified representations of a person's activities to help a person *see in the data* where they may need or want to do things differently for their health. Intriguingly, one is responding to the data picture of one's self rather than necessarily a perceived or felt experience.

Once one identifies something to do, such as going to the gym when a pattern shows one doesn't move, Hekler and colleagues' DIY'ing strategies (Hekler et al., 2013) help develop and assess

making the practice work in one's life, while Agapie et al.'s workout system crowdsources tailored versions of that activity, again telling someone what to do (Agapie et al., 2016), and Daskalova et al.'s work helps assess and tune it (Daskalova et al., 2017). In related, but more disease-oriented health care, Duan and colleague's (Duan et al., 2013) work uses formalized experimental N-of-1 designs to find triggers for negative effects in order to eliminate them. Work lead by Panday (Pandey et al., 2017) translates this methodology into a generalized architecture to carry out individually driven, but widely distributed formal N-of-1 studies within any area of human performance to answer "does X do Y"? There is a strong emphasis, among this self-experimentation side, for producing rigorous statistical estimates (Lee et al., 2017; Phatak et al., 2018; Kravitz et al., 2020)—but that level of rigor is not always appropriate, particularly for early practices when a person is in effect, simply trying to determine even if they want to try out a given health behavior, little own commit to it for the rest of their lives.

By looking at work in HCI in one exemplar health area, sleep self-study, we see these clusters more sharply. Here, most work has focused on blending *what to do* in terms of sleep practice *via* tracking and analysis of associated sleep data [SleepTight (Choe et al., 2015)], building peripheral awareness of a recommended sleep enhancing protocol [ShutEye (Bauer et al., 2012)], and making recommendations of how to refine a sleep protocol based on external expert data analysis/pattern matching [SleepCoacher (Daskalova et al., 2016)]. In other words, the person outsources what to do about their sleep to an expert, data-driven guidance system. The prescription seems personalized based on one's own data, and the success or failure is measured by adherence and improved performance based on measured data. In what is framed as "autoethnography" around sleep, Lockton and colleagues recently focused on participants designing probes that would let them reflect on aspects of sleep important to them, individually (Lockton et al., 2020). Here, the emphasis of the work is on reflection on current practice rather than changing it to sleep better. But even here, the majority of designs in the concluding "insights" section come back to how to *re-present* data already generated in sleep trackers, for instance, to look at patterns of sleep in different contexts. These approaches tend to assume these data are valuable, informative, sufficient, trustable. We let the data tell us how we are doing, and for many who adopt these approaches, these assessments are taken very seriously. Research around sleep tracking shows increased anxiety around sleep performance as a result is not uncommon (Baron et al., 2017). But research in sleep trackers also shows us how inaccurate they are (Ameen et al., 2019; Tuominen et al., 2019; Louzon et al., 2020). This is a conundrum for design, raised previously in a discussion about the "accuracy" of scales for supporting weight management (Kay et al., 2013): are we achieving what we hope to achieve for health with this data-led approach? Do we need to better qualify what the data presented is actually able to tell users about their experience, and their progress in a health journey?

Research in HCI has also shown that not everyone who might benefit from health support from interactive technology finds these quantification/tracking data-dominant approaches

¹<https://www.apple.com/uk/watchos/watchos-7/>

appealing (Epstein et al., 2016). This efficacy seems to be mostly for people already committed to exploring either their general health (Clawson et al., 2015) or to address a very specific health need (Karkar et al., 2016) and that even within that population, abandonment of trackers and self-tracking is high (Sullivan and Brown 2013) (Stawarz et al., 2015). Thus, there are at least two communities where the potential benefit of interactive technological interventions has not created benefit: those who find quantification a non-starter; and those who abandon it—without necessarily building a sustainable practice with it. There is also a related problem; even among ardent trackers, the way these technologies are designed sets up the very high risk that, if/when the technologies are removed, the desired change will also go away. This general pattern was shown in prior analogous work related to gym membership (Hekler et al., 2013). In particular, a previous physical activity intervention, which was shown to increase physical activity relative to a control condition in a randomized controlled trial (Buman et al., 2011), included the option of gym memberships, though not everyone in the study used the gym membership. While there were people who increased physical activity both using the gym and not using the gym, it was only the people who did not use the gym who maintained their physical activity when the intervention was over and the gym membership was taken away (Hekler, 2013). Returning to fitness programs and trackers, what happens on day 91, after the 90-days fitness follow along? This question of course reveals a concern dominant in this paper: what do we want our interactive health technologies to enable? To support building device reliance or independence? Another way of framing this question may be: does outsourcing to health technology make us more cyborg (Haraway, 1991) than human? Again, that continuum—raw to cyborg—is not meant as a value judgment, but as a way to be, perhaps, more deliberate and specific in our design choices.

For those who have abandoned or do not engage in data-dominant health support, we suggest that the tracking/quantification may not itself be the main issue for resistance, but rather the *outsourcing* of a practice to a quantifiable data driven approach, a point that is implied with Wolf and DeGroot's framing on personal science [note, Wolf was one of the originators of the notion of Quantified Self and, in this piece, he is signaling a shift away from the emphasis on quantification (Wolf and De Groot, 2021)]. The tools, in the end, can be perceived to be too brittle or narrow at this point in their evolution to be more generally useful. Outsourcing is not just about giving monitoring of biometrics over to devices; it also needs the data-driven system to tell one how well they are doing and, what to do. In the authors' work with athletes, it's not uncommon to hear of dissonance between an athlete's watch telling them about their recovery state and how they themselves think they are feeling. In the sleep trackers, it's not uncommon for people who wake feeling ok, to be frustrated and therefore not seeing their deep sleep change. We suggest it may be that the perceived quality of the assessed states or recommended practices are not sufficiently responsive for the cost of perpetual tracking. Periodicity, meso, macro, micro cycles familiar in sports science is also largely absent. An implicit quality of most health devices is that they have no time limits, the implicit design being, the

intervention will last forever. That too has been noted (Churchill and schraefel, 2015) as a frustrating assumption in such outsourced health tech design.

How Do You Feel How You Feel vs. What Does the Data Say?

There is a challenge with insourcing-oriented designs: how to support internalizing awareness and knowledge of personal state? When we rely on outsourcing, on data to tell us how we are performing, we do not have to determine how we feel; we do not have to focus on building the pathways between our own embodied sensors and the machines'. Building up these lived connections—to be able to answer “*how do you feel how you feel*” is non-trivial on at least two levels. First, some people have not, at least in recent memory, had the experience of, for example, a good night's rest: they wake every day under-slept, roused by an alarm, instead of waking on their own. If we do not know what it feels like to be well-rested, or we cannot perceive these differences in terms of affect (Zucker et al., 1997; Erbas et al., 2018) then we lack a yardstick both to judge actions against, and make decisions about the perceived cost. Without this internalized yardstick and sense of difference, any information provided on plausible difference (you'll be well-rested) is a purely intellectual value. Thus, the plausible benefits one would experience are not part of the calculations for choosing different health actions when insourcing is not present (Vandercammen et al., 2014). Second, while humans have incredibly sensitive dynamic sensing systems, some people do not have the connection between that sensory input and their experiences well-formulated. They may be *interoceptively* (Tsakiris and Critchley, 2016; Murphy et al., 2019) weak. That is, without the lived *allostatic* experience (Kleckner et al., 2017) that connects, for example, positive movement with good mood, or nutritious food with cognitive effectiveness, the person does not have, literally, a mental map to help navigate toward healthful practices and associated experiences. Those pathways have not been built. This is critical based on new theories around subjective states and our understanding on what we feel and why. For example, there is increasing evidence suggesting that emotions are not innate to our biology or brain but, instead, appear to be a psychological construction of the interaction between interoception and context and history (Lindquist et al., 2012). The implication of this is that, if one does not develop a robust practice of interoceptive awareness in context, one quite literally, will not be able to meaningfully feel (Lindquist et al., 2012). As classic work in neuroscience and social psychology illustrates, this is critically important as, there is increased recognition that it ultimately our experience of emotions that guide on in the actions we take (Spence, 1995; Haidt, 2012). Thus, training on how to feel, meaning how to meaningfully experience and interpret the signals that flag up our state across contexts, is a critically important and, as of yet, under-studied area of work for advancing health. This is the domain that we are exploring in the XB framework.

Heuristics as Complement to Habits

Just as computer science and medicine synergize around capture and use of quantified data for health interventions, psychology

and computer science in HCI meet at Behavior Change². A key embodiment for behavior change in HCI is *the habit*: identifying (Pierce et al., 2010; Oulasvirta et al., 2012), breaking (Pinder et al., 2018), and especially forming (Consolvo et al., 2008; Stawarz et al., 2015) habits are strong foci of that engagement. The focus on habits is understandable: while the space of sustaining a health practice is multifaceted, including “motives, self-regulation, resources (psychological and physical), habits, and environmental and social influences” (Kwasnicka et al., 2016), habits themselves are highly amenable to what computational devices can support: context triggered cues, just in time reminders and compliance tracking (Rahman et al., 2016; Lee et al., 2017). For context, in psychology, habits are described as behaviors that, once developed, are enacted automatically, that is without much conscious thought, and within specific contexts (Wood and Rünger, 2016). This context-based automation is valuable and important as habits can facilitate appropriate rhythms for achieving meaningful health targets. For instance, “at night while watching the TV, floss teeth”. The TV, the night, the floss container at hand, all contribute to the cues to trigger that practice. Such habits are extremely helpful: who wants to think deeply every day about the precise practice of dental hygiene? But what if the context changes? If the context switches, such as while traveling, or the floss runs out? Habits, by their very specificity, do not support adaptivity, yet many health contexts require a broader range to support the goals of a practice. For instance, if the intention is to maintain fitness, and a device only supports “go for a walk after breakfast” what happens when one is where that walk is not an option? Where are alternatives? To address these challenges, we propose expanding “habit” as a single design point, into a continuum that includes *heuristics*.

Heuristics, like habits, are also well known in HCI, specifically as a method of usability evaluation where a set of general processes or principles (the heuristics) are provided to be used by an evaluator inspecting an interaction for those processes, like “visibility of system status” and “recognition rather than recall” (Nielsen and Molich, 1990). Heuristics like habits can be internalized, but here rather than specific autonomous actions, they are integrated as functional principles to be instantiated based on personal knowledge and skills that guide a person’s practice in any given moment. As such, they are a useful mechanism for *insourcing* dynamic, responsive, healthful practices. For example, in strength training, there is a heuristic: when in doubt, choose compound movements. Such movements work most of the body and are therefore “go to” work for general physical preparedness (Verkhoshansky et al., 2009). As part of building an insourced practice, one then builds up experience with a lexicon of movements that will support that heuristic: pull ups, squats, rows, deadlifts and so on. Similarly, survivalist Bear Grylls is famous for his heuristic: “Please Remember What’s First: protection, rescue, water, food” (Kearney, 2018). Success with this life-saving mnemonic

depends however, on knowing what protection is in the desert vs. the jungle. Likewise, one may need *practice* with the *skills* to make a fire to melt snow for water in the tundra, ward of other creatures, or be noticed for rescue.

Deliberate Choice in Habit-Heuristic Continuum

In weighing design choices for health at one side of this continuum, habit formation requires *intention* (Prochaska and Velicer, 1997; Hashemzadeh et al., 2019), what we might frame as insourcing of motivation, to put the effort into repeating the task sufficiently for it to become automated/habituated. Assuming that motivation exists, the triggers for firing habits as per the references above can be supported by technology (as per Stawarz et al., 2015), and thus their development may be effectively outsourced. However, like the data-prescriptive outsourced approaches discussed above, habits do not deliberately build knowledge skills or practice for executing options either. Heuristics do focus on building those skills. While heuristics are more adaptive, as more context-agnostic than habits, they also require potentially more resources to insource in order to develop the variety of experiences and expertizes to be independent of external guides. Thus, the habit-heuristic continuum opens up a way to let designers ask more explicitly both what they are trying to instantiate in a person, and what the costs/trade-offs of either approach may be.

Behavior Change vs. Knowledge Skills and Practice in Continuum

Another difference between habits and heuristics that interactive health tech may leverage is that, while habits are strongly associated with behavior change, heuristics are not. Heuristics are more associated with knowledge building and application: to have a general rule on how to approach a problem is different than a need to change in a very particular manner. While the ends may often be similar, the pathways, and thus design processes open to explore, are different, offering HCI additional options for designing toward successful health engagement.

XB—EXPERIMENT IN A BOX

Based on the above related work, we see considerable open space to explore the design and development of *insourcing health heuristics* to support building knowledge, skills and practice for personal health resilience. The following work presents XB as just such an exploration of this space. In the next sections we present an overview of the XB approach, several exploratory studies to define and confirm its key features, and then present an XB framework to enable general XB development by others that incorporates insights from these exploratory studies.

The XB approach provides a structure for translating a general health heuristic into a personal, practicable, robust health heuristic. For many, generic health heuristics are known, such as sleep hygiene rules, but they can also be hard for people to apply in their own lives. The self-determination theory postulates that to make the effort—to have the motivation—to engage in a health practice, a person’s autonomy, competency and

²In just the past ten years, 6481 papers in the ACM Digital Library include “behavior change” in the title, most in HCI (sigCHI) associated venues.

relatedness must be supported, and when they are, increased feelings of vitality ensue (Stevens et al., 2016). Therefore, with the XB approach we can create simple, short-duration experiences, “experiments” that empower people to test out a health heuristic and to build competency through the testing process. By experiment, we are explicitly using the word in a colloquial sense, which we will unpack more momentarily. Experiments, in a colloquial sense, are what we want to inspire. Through experiments, one can not only learn a potentially new health skill; they also learn the meta skill of how to explore and find strategies for health that can work for them (i.e., experimentation and trial and error on self). Based on this, we see the XB model as a skills-building approach for supporting self-study practice around health heuristics that cultivates both self-regulatory skills in the form of self-study and skills in doing specific actions that produce desired across contexts. It is in line with the personal science framework (Wolf and De Groot, 2021), but simplified in terms of providing initial heuristics to work off of, compared to more of a blank slate implied in the personal science framework. We help participants share their experiences both through talking about their experiences as well as sharing anonymized data of their experiences for open science (Bisolo et al., 2014).

To reinforce this exploratory approach, we frame each instantiation of the XB model as an “experiment in a box”, an XB. By experiment, we are explicitly *not* referring to the current health science use of that word, which connotes the use of randomization to enable a robust causal inference. As said, we use “experiment” in a colloquial sense as the colloquial concept aligns with the experience we offer. This framing as “experiment” has several SDT related advantages: it reinforces the autonomy of the Experimenter as the one in control, not the one being controlled or told what to do. The nature of an experiment is also time-limited, thus enabling a person to “try it out before committing” to a particular health protocol. That approach is in contrast to the dominant paradigm that assumes one will choose and follow a strategy forever. The “box” part of the Experiment in a Box is the guidance the XB provides to test and evaluate the heuristic and to translate it in the process from a general heuristic to personalized health *knowledge skills and practice* (KSP). The box is where the key distinction from the more general personal science framework occurs.

KSP is important in the XB model. Knowledge is prior information learned by others or oneself relative to the relationship between actions and outcomes, both positive and negative. Skills are strategies, approaches, and actions one actively engages in to achieve desired goals. These skills can be far ranging from the concrete development of one’s physical ability to do an action (e.g., shoot a basketball into a hoop) to one’s self-regulatory capacity to define, enact, monitor, and adjust one’s actions in relation to a desired state. Practice has two associations, as pointed to in the introduction. The first is simply the repetitions necessary to be able to use a skill, apply knowledge effectively. The second is the incorporation of these skills *as* something that is deliberately, actively and consciously part of one’s life.

In the first case practice involves cultivating personal knowledge, which includes both that which can be expressed, and also what may be tacit knowledge—that is, “known” by the

person but may not be conveyable *via* language (e.g., how to kick a football or play a chord on a guitar; building one’s capacity to know how to adapt appropriately in various contexts). For example, one general health heuristic for a healthy sleep practice might encapsulates actionable knowledge as: “ensure that the environment where you sleep is dark when sleeping”. The skill is the ability to execute that knowledge in diverse contexts. It might include something as simple as ensuring lights are off or blocking light leaks from windows. The more diverse the sleep contexts one experiences where they solved the dark room challenge, the broader the options are to draw upon to execute this heuristic across contexts. By extension, a person’s robust practice across contexts leads to enacting practice agility—a significant component for “resilient” (Feldman, 2020) health practices. As a general health heuristic becomes tested and refined by/for each person across contexts, it becomes an increasingly more personal and robust health heuristic, hence a personal health heuristic. For example, one may find that 1) dark rooms improve their sleep and 2) their best path to a dark room at home is a sleeping mask, but on the road, traveling with a clothespin/peg to shutter a hotel curtain, is essential. XBs are designed to help people develop this KSP experience/expertise.

Practice in the second sense is also a way of framing an ongoing engagement with knowledge and skill work. One may have a professional practice, a spiritual practice, a physical practice. The concept of practice in this way situates the interaction as deliberate, intentional, and continuous. It also inherits, at least in part, the sense of practice as part of continual skill refinement. This sense of practice further differentiates our focus on heuristics from habits. A habit’s great strength is to be able to “set it and forget it” like “brush your teeth every day on waking.” A heuristic-as-template on the other hand helps guide and instantiate knowledge, and apply skills, in diverse contexts. Thus we practice skills like a tennis serve, at first to learn the technique and build the coordination until the skill is automated (Luft and Buitrago, 2005)—where it actually changes where that pattern sits in our brain. After this, we practice the skill literally to keep those neural pathways cleared and strong (Olszewska et al., 2021). But the practice of the practice is to deliberately work to add nuance to the skill, to operate better, faster, across contexts. Such ongoing refinement is necessary, for example, around how we eat. What diet supports muscle building in year one needs to be refined in years two, as the body adapts, circumstances change. Thus, one’s physical practice is both regular, one might say habituated, *but* critically, is also conscious, deliberately open to refinement to support continual tuning. We will see this latter sense of a practice in KSP as part of the *insourcing* we have been developing the XB framework to help establish and support.

EXPERIMENT IN A BOX EXPLORATORY STUDIES: TOWARD THE EXPERIMENT IN A BOX FRAMEWORK

The above section lays out our general thinking that guided of how we designed our initial explorations into XB. In this section

we present how we iteratively refined and tested our approach over three studies. We had several goals in these studies. A key goal is to see that health skills tested in an XB are effectively insourced. We define that in terms of a recognition that the specific health skill a person experimented is used over time. Further, we also sought to see if people insourced the more general skill of self-experimentation. Do people report continuing to try out different generic heuristics they hear of in their lives, as is suggested they do in the XB framework? In addition to the behaviors we were seeking to observe, we also wanted to gain an insight on perceived benefit. As our focus is on insourcing and interoceptive awareness, we asked an intentionally broad question: does a person feel better? Further, do they feel more capable in managing their wellbeing?

Our study approach, which involved three studies over time with increasing levels of rigor, was grounded in the classic scientific logic of assumption articulation and testing (i.e., iteration) coupled with triangulation toward consilience. Specifically, based on the stage of this approach, which as we argue, is still very nascent, we explicitly started highly exploratory in our approach, to enable experience to guide our next steps. As we learned from experience, we refined our testing protocols. While no one study of this sequence provides definitive evidence of utility of the XB approach, the similar patterns we saw across different populations, ways of measuring, and ways of supporting an XB experience (i.e., consilience), we contend, is suggestive of the value of further study and exploration into the XB framework.

In addition, our iterative study approach was inspired by the ORBIT model (Czajkowski et al., 2015) for behavioral intervention development. In particular, study one and two were both variations of a proof-of-concept trial, with study one used to help refine XB (i.e., the end of phase I of the ORBIT model), and then to gather evidence of real-world impact relevant to the outcome of interest in study I, which is the start of phase II of the ORBIT model, focused on justifying the need for a more rigorous clinical trial. And even in our third study, we used more of a proof-of-concept formalism for evaluation, which explicit does not use statistical analyses between groups. The reason why we used iterative proof-of-concept studies is because, both, this is the stage of development we are at (define, refine, test for a meaningful signal), and because this approach is increasingly recognized as a more appropriate method for exploration than the increasingly debunked strategy of running underpowered randomized controlled trials and conducting corresponding “limited efficacy” analyses of between-group differences. Indeed, under-powered pilot-efficacy-trials explicitly are increasingly being shown to produce poor evidence that actually stymies scientific progress (Freedland, 2020a; Freedland, 2020b). The short reason is that, with such under-powered trials, any effect size estimates gleaned from such trials is highly untrustworthy and, thus, largely not interpretable. To avoid these traps, a series of proof-of-concept studies, which we conducted, fits with emerging best practices for work focused on defining, refining, and testing for a signal of an intervention. Further, when looked at together, they provide increased confidence, *via* the notion of triangulation and consilience,

that the XB approach has been specified with sufficient rigor to enable limited replicability across populations. As illustrated in ORBIT, this does not mean that any claims of efficacy or effectiveness relative to some meaningful comparator can be made at that time, but, again, that was not our purpose. Our purpose was to define, refine, and test for signal (i.e., phase I and the start of Phase II of ORBIT). More details on each proof-of-concept study variation below.

Study One presents a low-fi exploratory study with a 12-member participant group to see simply *if* these short experiments were positively received, experienced as useful, and in particular, lead to lasting use after the study intervention ended. The results were overall positive, and we present a set of design insights from that study that formed the basis of a more formal interrogation in *Study Two*.

In *Study Two*, we translated the insights from study one into a stand-alone application to be used by individuals, independent of a group context. Using a within-subjects design, we sought to assess if such an implementation was sufficiently robust to create a positive effect during the study, and could likewise instill lasting knowledge skills and practice, as indicative of the creation of a personal health heuristic, beyond the end of the main study.

The outcomes of the study were overall positive. In *Study Three*, therefore, we sought to test our assertions on the key elements of XB—the framing of an experiment; the support to regularly evaluate and reflect on its utility—actually enabled XB success beyond simply making good practice guidance readily accessible. To this end, we ran a formative between-subjects randomized study to test our hypothesis that an XB approach would be more effective at promoting engagement with health practices compared to a version that mirrors standard mechanisms of offering health support. For our test we reused the SleepBetter XB from Study Two and created a similar app that offers knowledge (i.e., information about sleep) and details sleep health behavior skills (i.e., sleep hygiene), and invites a person to try any one of these practices for a week but does not provide the self-regulatory skills exploration operationalized by the XB model, nor a structure that guides experimenting with one’s practice. While we imagined that the more supportive approach would be better than the current standard for sharing health practice building, the strength of the responses between the two conditions underpinned the value of the XB.

Study One: Exploring Experiments for Sustainable Health Practices

In Study One, we worked with 12 participants from a national advertising organization over 6 weeks. As creativity is an essential quality for advertisers, the motivation for participants was their interest in improving their health to improve their creativity. As measuring creativity directly is notoriously messy (Thys et al., 2014), we drew on related work that shows both: 1) how cognitive executive functioning areas of the brain map to associated areas for creativity; and 2) that fitness practices enhance those areas of the brain (schraefel, 2014). From that, we postulate the use of validated cognitive executive functioning tasks to demonstrate quantitative benefit, along with participants’ qualitative

TABLE 1 | Study One in5 experiments.

Yellow box: move	Every hour at work, stand up, go for a one hundred pace walk away from and back to your desk; use the stairs or walk to achieve a pace to get your heart rate up “a bit.”
Green box: eat	Any time you eat, have anything you wish, but have a green veg and some protein before you have anything else
Red box: engage	Twice a day, ask how someone else is, and for 3 min, listen, without verbal interruption
Blue box: cogitate	Learn a new skill: one you can practice daily, many times, for a total of at least 20 min
Black box: sleep	For 1 week, with manager’s permission: no alarms. Wake up when you wake up. Your action: try to get to sleep 2 h before midnight

self-reports of their experience. Participants carried out several of these tests prior to beginning the formal study and after each week. These included strop tests, tapping tests, memory tests (Chan et al., 2008).

In this exploratory study, we focused on offering participants a weekly “taster” of five fundamental health practices for wellbeing we called the “in five” of move, eat, engage, cogitate, and sleep, detailed in (schraefel, 2019). In the six-week engagement, week one was a preparatory week to learn about the study and how it would run each week. In each of the 5 weeks following, the group ran an experiment on one of the in five. At the start of each week, the team manager sent around a document we co-wrote, with descriptions of the experiments and instructions for the protocol, asking people to track their practice/experiences at least once a day. We did not specify how to do this tracking; we were interested in what people would find important to meet that observation requirement. The manager was very good about sending out reminders every other day to make sure people had logged something about their experience for that study. To align these experiments as much as possible with the group’s motivation around creativity and design, we also named these experiments with colors rather health concepts. For example, the Black Box was the Sleep protocol; the Green box was for Eat, the Blue Box was for Cogitate, Red Box for Engagement and Yellow was for Move. **Table 1** lists each of the experiments, and their associated protocols.

In this first study, the experiments for each week were predefined, but we emphasized choice in the variety of ways the heuristic could be implemented. For example, in the Green Box, the heuristic was “up your green and red” green being green vegetables; red being any protein. The experimental protocol was: “eat any time you wish, as many times in a day as you wish, but any time you eat, for the five-day work week, have some green veg and some kind of protein before you eat anything else, any time you eat”. As a group, prior to the Green Box start, we reviewed what constituted a “green veg” and what a protein was. To reduce any sense of sacrifice around eating in the experiment, the approach was also deliberately additive rather than subtractive: “as long as a green and protein are present, add anything else after that; and also have the green and protein before eating anything else you add”. We also acknowledged that this was not likely how people would continue to eat after the experiment, but that as an experiment this was to ensure dose effect in the given time.

At the end of each week, we asked participants to capture: 1) what did you learn/experience? 2) based on the effect experienced, how would you turn this into a part of your

life—how would you create a personal health practice (a personal health heuristic) from this experiment? 3) what are any questions you have right now about implementing that practice? In our Black Box for sleep experiment, the kinds of questions we would get were: “Sleeping till I wake feels great; how do I do this on days when I want to go out with my friends?” This led us to discuss sleep debt, recovery, and also balancing values like social interaction for wellbeing with rest and recovery. In the Yellow Box, we learned that people were reluctant to be seen leaving their desks so frequently lest their colleagues perceived them to be slacking off. This response offered a powerful insight for managers in terms of the interventions needed to support movement as a cultural practice to be encouraged.

At the conclusion of each week’s experiment, we met online *via* Skype to reflect together on outcomes from that week and to prep for the week coming up. In-between weekly meetings, we used an open-source online bulletin board to share observations, respond to questions, and add reminder requests to record experiences. As noted, participants were also invited to re-run their creativity assessments once a week, post each experiment.

In this study, we did not ask participants for their logs or their creativity assessment data: that we agreed at the outset was personal, and we were keen for as many people on the team to feel as safe as possible working with us as co-explorers, not subjects. We agreed on sharing material on the forum with all of us in our weekly meetings—including the three questions asked each week, above. Our final interviews would be used for our analysis. To this end we asked participants to share as much of their insights as they felt comfortable, to help inform the design process.

Design Insights From Study One

Overall, the consensus was that the process was interesting and surprising in positive ways. The group even made a 3-min video to reflect on the experience (Ogilvy Consulting UK, 2014). From the forum posts and our weekly interviews, we gathered the following key points around factors affecting the utility of the XB approach:

- 1) *Small repeatable doable doses; big effects.* In developing the experiments, we were very careful to ensure that an effect could be felt within the week of its practice. We drew on our own professional backgrounds in health and sports science, consultations with colleagues who coach health practices professionally, and associated literature reviews to develop these protocols. Despite this preparation, we were surprised by

the reported size of the experience based on the conservative nature of the protocol. For example, in the Green Box, we were particularly concerned that any food intervention would have a noticeable effect within a week, as the usual report in the literature had been 14 days to show measurable results. It seems experiential results may be sooner: each participant reported even before the week was up about changes in energy; about the surprise as they thought it would be hard to get a green and a protein in all the time, but how easy it actually was. It seems also that dose size afforded by an experiment is also valuable: that each time a person ate, they focused on the green/protein combination. This meant they were paying attention to this practice and its effect—both on immediate eating experience but also in reflection—having many data points to draw upon by the end of that week. Thus, short, simple, multiple doses that are readily achievable (with preparation) and that will also create an experienceable effect within the period seem to be key.

- 2) *Beyond knowing: creating space to do the practice.* Participants also commented on the importance of direct experience of a practice they “knew” of but had not directly tried. For example, in our Black Box for sleep week, we worked with the Group’s management to request permission to let participants test the heuristic “wake up when you wake up” and thus to sleep as long as they needed, and to come into the office when they were up and ready, without penalty. Intriguingly, it took more effort from the manager to confirm with participants that this was OK, than it was to get that support in the first place. Once they had the space to embrace what became known as “kill the alarm”, they did. Again, they reported value around being asked to actually “test” the concept in the experiment. As one participant reported, “I’d always heard that 8 h of sleep was beneficial, but until I actually tried it, I had no idea how much of a difference it made.”
- 3) *Unanticipated benefits*—While some people expected they may have more energy or gain more motivation to stick with a health practice, they were surprised to find how quickly their mood also seemed to improve; and how easy these practices were to make a regular practice. One person’s surprise was in their colleagues, especially after the Red Box on Engagement where they were asked to take 3 min every day to listen to a colleague, without interruption. “I used to hate that person, but after listening to her and others, it was amazing, she became so much nicer; she stopped bothering me.” There may be an opportunity here for design to explore more deliberately unanticipated positive side effects associated with experimental experiences.
- 4) *Overall effects appeared to be cumulative*—Each week people reported positive experiences; after week three, most were reporting they were starting to notice positive differences in their mood, sense of wellbeing, energy, creativity and for some, their weight. This progressive experience across experiments may be another side effect that can be explored more deliberately in design: to cue up that there may not only be benefits from the time given to one experiment, but in spending this period—in this case six consecutive weeks—on health practices; there may be additional benefit just in the exploration of health for those several weeks. This experience in itself of having the permission space and support to focus on health KSP—irrespective of experimental outcomes—may foster a foundation for lasting health engagement.
- 5) *Short exposure may lead to lasting practice*—In follow ups—3, 8 and 13 months after the study, participants were still using skills they had learned—translating them into practices that were working for them. For example, one participant shared that they attempt to get full nights’ sleep for them at least four nights a week so they can have “guilt-free socializing time” with their colleagues on weekends. The participant claimed the XB experience helped them explore and build a sleep-recovery approach to support their wellbeing. This approach was not taught by the Black Box, but the Experiment aspect of the study itself encouraged the person to keep testing practices to work for them.
- 6) *Skills Debugging*—Related to ongoing use of both skills and experimental “test the approach”, we found that, without prompting, participants used the skills from an XB to debug and retest their practices. One participant reaffirmed her initial results with one experiment that for her, a diet high in greens with sufficient protein feels good on numerous levels. She told us: “My parents visited for a week, and I just stopped doing Green and Red (a green and a protein with each feeding), and I ate what I’m like when I’m at home. Lots of bread, lots of pasta—no greens, very little red. I felt stressed; I reverted to type. When they left, I felt crappy and bloated and no energy. I decided to go back to red and green; the weight I put on had gone and I feel so much better. Maybe the stress is gone because the visit is over, but I think it’s a lot about the food.”
- 7) *Preparation* This result is inferred by participants’ discussions about their practice: in many cases the way they described carrying out a practice drew from the preparation for it in the week preceding it. For the Blue Box, the experiment was “learn new skills you can practice daily, many times, for a total of at least 20 min”. Here, we heard about how people had put together tools they would need, for orienteering (“I’m directionally impaired”) or sketching materials and where tools would need to be when setting up their “lab box” for the experiment. Likewise, for the Green Box, how they would have containers and prepare vegetables for the week to come again to have their lab ready to go. Preparation also helped illuminate challenges that we had not considered and needed to address on the design side. In the Yellow Box, we asked people to move away from their desk twice an hour and to keep moving for a few minutes each time. Operationalizing this was not a problem; fear about how their colleagues not in the study would feel about them walking away so frequently was a challenge we were able to mitigate by working with management ahead of that box’s week.
- 8) *Diversity/Choice* In this study, the diversity of ways to explore health was also seen as a benefit; it let us convey how health is dynamic and that there are multiple options within multiple paths to approach it. Enabling people to determine how they

TABLE 2 | Sleep experiments.

Factor	Key	Experiments
LIGHT	L1	Increase bright light exposure during the day
	L2	Wear glasses that block blue light during the night
	L3	Turn off any bright lights within 2 h of going to sleep (such as the TV/the computer etc.).
CAFFEINE	C1	Do not drink caffeine within 6 h of going to sleep
	C2	Limit yourself to 4 cups of coffee per day/10 cans of soda/2 energy drinks
	C3	Do not drink caffeine on empty stomach
SLEEP SCHEDULE	S1	Usually get up at the same time everyday, even on weekends/vacations
	S2	Sleep no less than 7 h per night
	S3	Do not go to bed unless you are relaxed/tired. If you are not, relax with a bath/by reading a book before going to bed etc.
	S4	Go to bed at 22:30 PM the latest

would choose to implement a heuristic was also well received: “We could eat whatever green we wished—I learned about so many veg I’d never tried—it was great.”; “I wanted to learn how to play guitar—I didn’t know that was good for my brain, too.”; “after moving and sleeping, I just felt so much more creative.”

Study One Summary

In sum, our work in this first exploration showed us that guided exploration of health-KSP in even short periods provided a basis for participants to build their own health heuristics. These short, focused activities led to enduring use without any further technical support.

Study Two: SleepBetter—Many Paths to Health

Based on the richness of our results from our first study, our goal in this phase of development was to explore how we could translate that lo-tech, guided XB approach into a stand-alone digital intervention that could potentially reach and benefit more people. There was also growing concern at our University for students’ wellbeing, and the correlation of poor sleep with mental health challenges reported to be on the rise in this group (Milojevich and Lukowski, 2016; Dinis and Bragança, 2018). Therefore, we decided to explore how an XB could be redeployed to focus on one health attribute, in this case Sleep, while still building a sense of exploration and skills development. As well, we saw this XB implementation as an opportunity to complement the predominantly outsourcing-oriented work around sleep in HCI, as overviewed above, with a study of an insourcing approach.

In this single-focus XB, we were also able to shorten the engagement with the XB from 5 weeks across five topics to 15 days on one, where participants carried out 3, 5-days sleep experiment cycles. We facilitated agency by offering participants a suite of ten experiments to choose from to test sleeping better (Table 2). Further, participants could choose to carry out the same experiment for each cycle or choose a different one, from the set of ten.

In research, self-perceived sleep quality is considered to be a more important marker of success rather than duration (quantity) (Choe et al., 2015a) alone (Mary et al., 2013;

Mander et al., 2017; Manzar et al., 2018). Thus, our heuristics/experiments focused on practices that were assessed on a qualitative and experiential basis, rather than the currently dominant and possibly inaccurate (Ameen et al., 2019; Tuominen et al., 2019; Louzon et al., 2020) tracking approach of hours slept, number of interruptions, time assumed to be in a particular sleep state and so on. Our focus was: after trying this heuristic, how do you feel?

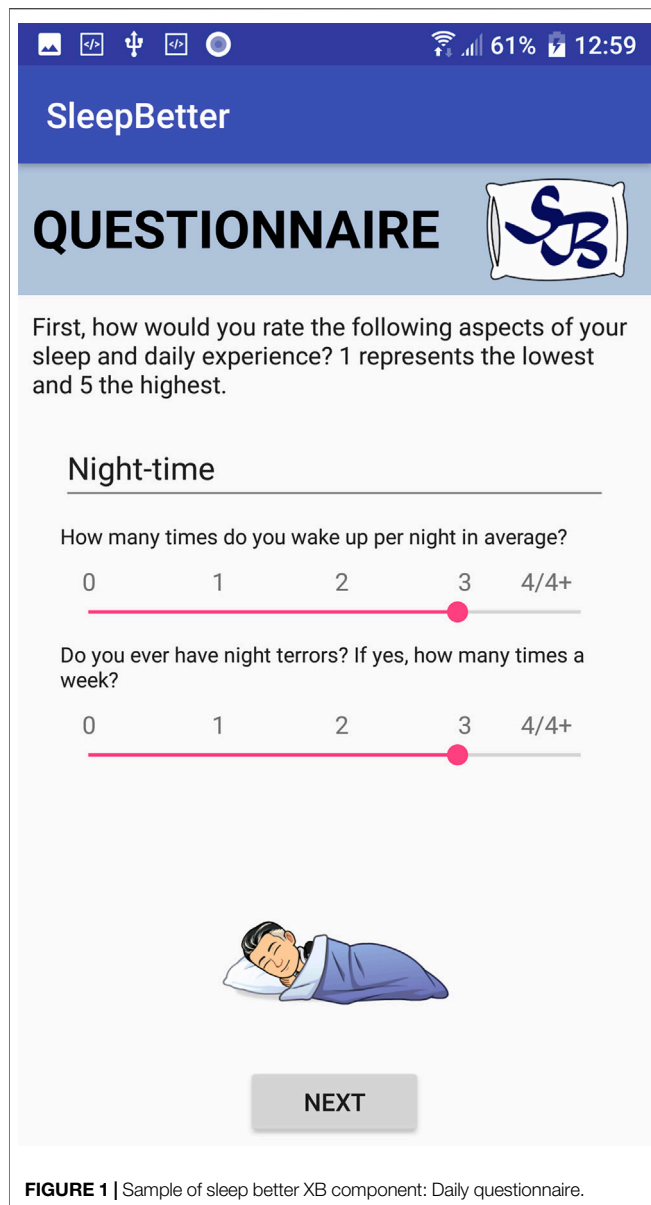
Apparatus: SleepBetter Experiment in a Box App

We developed SleepBetter, an Android OS smartphone XB application (Figures 1–3). The application includes 1) an experiment selection area; 2) self-reflection aids 3) a FAQ to provide information about the rationale/science behind each experimental protocol.

The application provides ten experiments for sleep improvement (also known as “sleep hygiene”) listed in Table 2. Each experiment is grounded in sleep research drawing on: guidelines from the National Health Service (NHS) United Kingdom (Sherwin et al., 2016) the Mayo Clinic, United States (Ogilvy Consulting UK, 2014), and related research (Bechara et al., 2003; Espie et al., 2014; Schlarb et al., 2015; Agapie et al., 2016; Daskalova et al., 2016). We also reviewed this set with sleep researchers.

Each day at the same time, participants were asked to respond to a short questionnaire (Figure 1) about their sleep experience the previous night, including: how they felt when they woke up, how they felt before going to sleep, their mood, and their perceived concentration. Each day they were also asked, summatively, “Do you feel better or worse than yesterday?” (Figure 2). Each question was set on a five-point Likert scale. Figures 1, 2 show examples of these representations.

A free-form reflection diary supported text entry for any annotations participants wished to make each day. Drawing on Locke and Latham’s approach to goal setting for motivation (Loke and Schiphorst, 2018) we called this space a Goal Diary so that participants could use this space to note a desired outcome from each experiment and reflect on effect. Participants could see graphs of daily progress based on their questionnaire answers; they also had a calendar (Figure 3) with circles around dates to show completed experiments. Circles were also colour-coded against a selected Mood state: greens (better); reds (worse).



SleepBetter

QUESTIONNAIRE

First, how would you rate the following aspects of your sleep and daily experience? 1 represents the lowest and 5 the highest.

Night-time

How many times do you wake up per night in average?

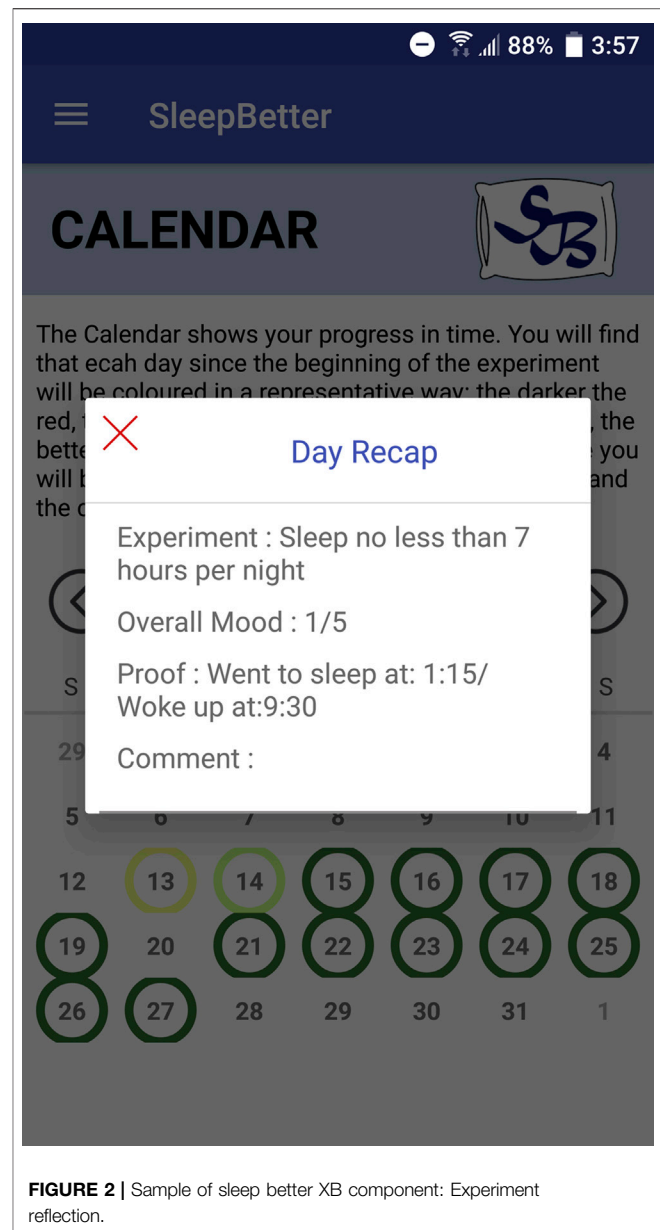
0 1 2 3 4/4+

Do you ever have night terrors? If yes, how many times a week?

0 1 2 3 4/4+

NEXT

FIGURE 1 | Sample of sleep better XB component: Daily questionnaire.



SleepBetter

CALENDAR

The Calendar shows your progress in time. You will find that each day since the beginning of the experiment will be coloured in a representative way: the darker the red, the better the mood you will be and the more the...

Day Recap

Experiment : Sleep no less than 7 hours per night

Overall Mood : 1/5

Proof : Went to sleep at: 1:15/ Woke up at:9:30

Comment :

FIGURE 2 | Sample of sleep better XB component: Experiment reflection.

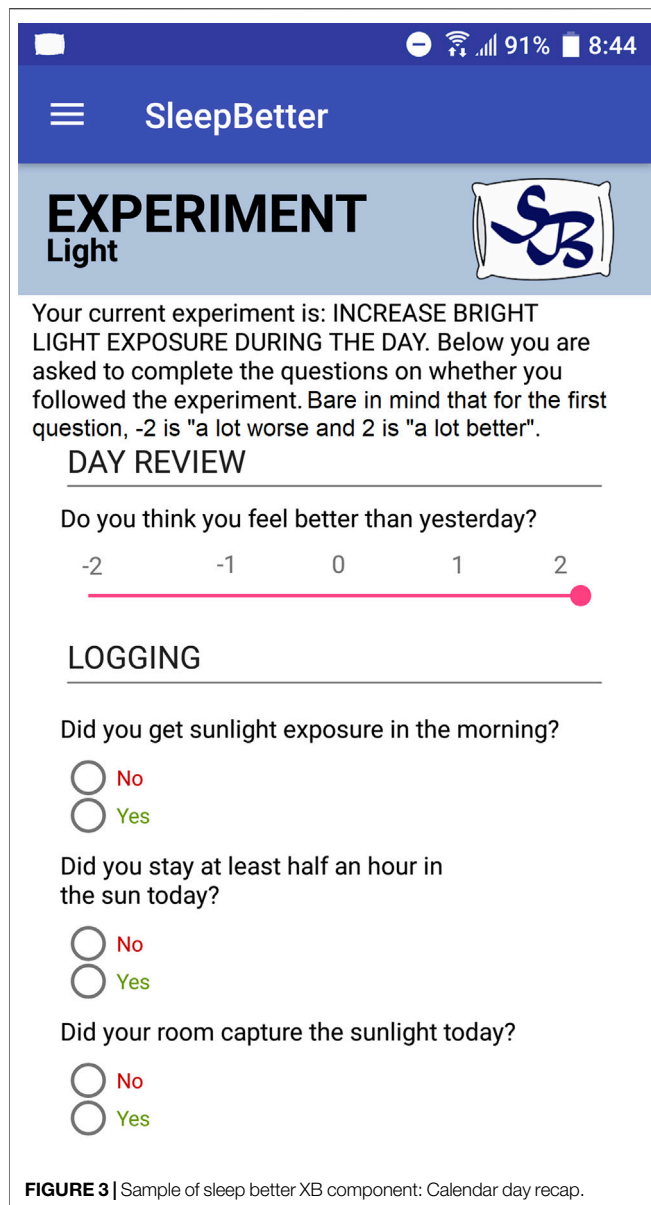
Participants

As noted, with our focus on university students' wellbeing, our inclusion criteria were for university students. To connect with this group, we used social media to recruit a convenience sample of the 25 final participants. The mean age was 21.8 (SD = 1.28). All participants identified sleep as an interest for themselves and agreed they wished to participate in the study either to learn more about sleep practices or because they wanted to improve their own sleep. A limitation of the study group is that, in this convenience sample, all participants identified as male.

Procedures

This is a within-participants study with three Phases: A, B, C. *Phase A* includes initial sign up, consent gathering, apparatus set up and baselining questionnaire that took place over a 6-day period. *Phase B* is the XB intervention itself, which lasted for

15 days and was broken into three, 5-days "experiment" phases. At the end of phase B, one-on-one tape-recorded interviews took place, pending participant schedules, within 10 days after the intervention. *Phase C* is a post-intervention survey covering self-report of wellbeing and sleep practices as well as reflections on the XB experience 7 months after Phase B, when all use of the application had ceased for 6 months. We also note that out of the 25 participants who completed the study, we only contacted 16 for Phase C who deliberately stated (in Phase B) that they agree to being contacted with a follow-up at a later date. On completion of Phase A, Phase B started. *Via* the app, participants were presented with the list of experiments from **Table 2** and were informed they could choose any experiment from the set but would not be able to change it once selected, until the 5-day period for each experiment was complete. To reduce the burden



SleepBetter

EXPERIMENT Light

Your current experiment is: INCREASE BRIGHT LIGHT EXPOSURE DURING THE DAY. Below you are asked to complete the questions on whether you followed the experiment. Bare in mind that for the first question, -2 is "a lot worse and 2 is "a lot better".

DAY REVIEW

Do you think you feel better than yesterday?

-2 -1 0 1 2

LOGGING

Did you get sunlight exposure in the morning?

☐ No ☒ Yes

Did you stay at least half an hour in the sun today?

☐ No ☒ Yes

Did your room capture the sunlight today?

☐ No ☒ Yes

FIGURE 3 | Sample of sleep better XB component: Calendar day recap.

of having to remember to engage with the app, the app sent reminders to complete a questionnaire for the day. The reminder encouraged exploration of the FAQ for questions about the experiments in general or sleep in particular.

Measurement

The main measurements during Phase B were the *daily questionnaires* which had two parts: protocol questions and state questions. The protocol questions were specific to each experiment. For example, for experiment C1 (**Table 2**) the questions included: "What time was your last coffee?" and "What time did you go to bed?" State questions focus on general sleep performance, such as: ease of falling asleep, sleepiness or alertness during the day, factors in mood, appetite and concentration ratings—all factors affected by sleep. As this study is exploratory rather than confirmatory,

we drew on questions where the effect of a sleep intervention has been documented within days. These questions were drawn from established sleep assessments (Espie et al., 2014; Ibáñez et al., 2018) to map to the brevity of our intervention period.

Changing Their Experiments

After each of the three 5-days experiments, participants were offered the opportunity to choose a new experiment or re-run their current experiment. They were also given an additional questionnaire on day five. This interaction explored:

- 1) What attracted them to the experiment they just completed;
- 2) Their experience, prompting them to reflect on its perceived effectiveness in terms of how they felt—and if they were changing or sticking with their current experiment for the next iteration;
- 3) What factors seemed interesting to them about staying with or changing experiments.

We emphasize that our goal in analyzing the data we captured from each of these questions was not to be confirmative, per se, but exploratory, to see the ways in which our design approach supported our aspirations to support connecting brief practice with a felt sense of benefit. To that end, we present findings and discussion, below.

Findings

Thirty five participants were initially recruited, out of which: 25 completed the study, four dropped out after signing up when they could not run the app on their particular phone's OS version; two reported that due to circumstances changing they could not participate; two gave no reason, and two dropped out when they changed phones during week one of the study. In sum, eight people dropped out during the pre-study set up phase, and two within the first 2 days of the study, having effectively null engagement with the intervention. As such, they have not been included in the reported data set. The overall change or benefit of the XB can be assessed in two ways: 1) as a perceived effect from pre-experiment assessment (what we call here Phase A) to post-experiment assessment (Phase B); and 2) by uptake and ongoing use of protocols over time, post-experiment (Phase C).

Phase A to B Effect

There was a statistically observed difference between A to B. Based on 13 questions used in sleep study assessments about perceived personal state and sleep quality, with each question based on a five-point score, where lower score was better than a higher score for a maximum score of four per state (**Figure 4**, Y axis), the Phase A results were 33.4, with SD = 6.52. The mean of the final questionnaire at the end of the study is 25.44 with a SD = 7.04, $t = 5.452$ (with $p < 0.001$). Cohen's d formula to evaluate an appropriate effect size for our means is 1.17, thus the measured effect for the difference between the means is strong. Results are presented in **Table 3**, where questions with a significant effect are highlighted.

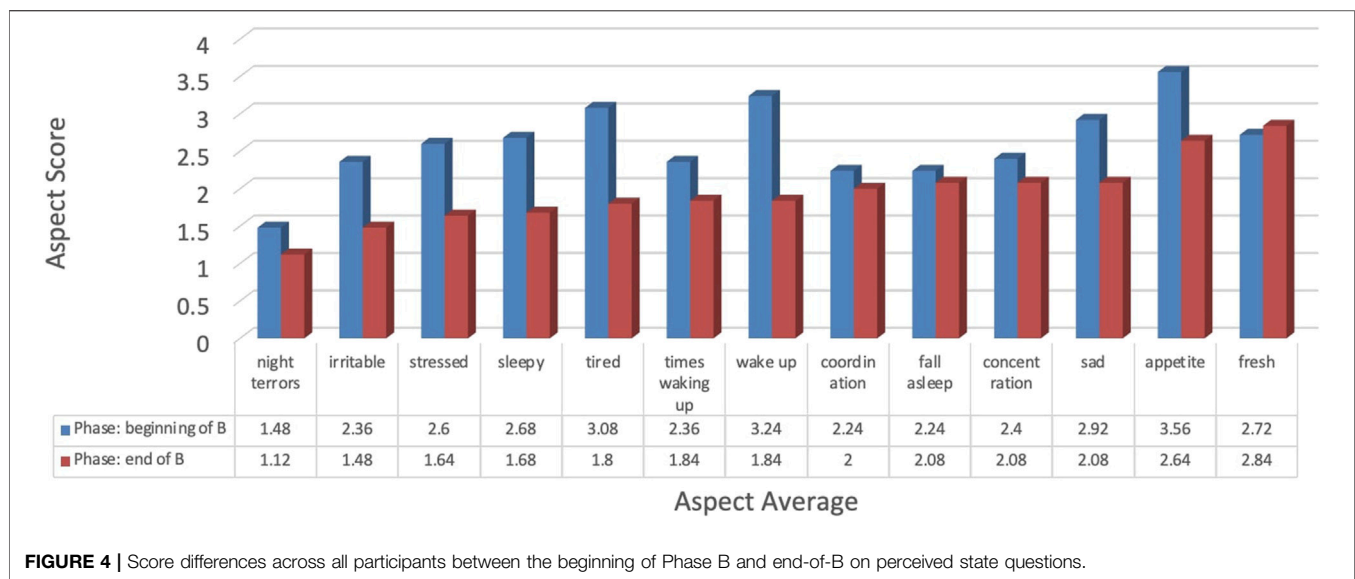


FIGURE 4 | Score differences across all participants between the beginning of Phase B and end-of-B on perceived state questions.

Looking within these measures, we see significant positive changes in sleep quality attributes: nightmares/night terrors were reduced. Mood categories of perceived tiredness, sleepiness, stress and irritability improved, as did appetite (cravings are reduced). The other categories: number of times waking up, ease of falling asleep, concentration and coordination changed positively but not significantly. The only slight negative change, though not significant, was feeling fresh when waking up (Table 3), which may be due to this being one of the first times people were focusing on this factor regularly.

Phase C Effect

Of the 25 people who completed the study, 7 months later, 16 agreed to complete a follow-up survey of their experience with the SleepBetter XB, and how this experience may have continued to affect their sleep practices. This questionnaire was entirely self-report, and so suffers from the known issues of overly optimistic assessment of performance. Even given that, we were surprised by the degree of at least reported persistence of practice. Comparing their sleep quality between before the study and after, 75% rated it now as “good” and 19% at “very good”, well up from their Phase A baselines.

In one question we relisted Table 2 experiments and their “tips”, like “getting at least 7 h of sleep a night”. All 16 respondents reported remembering all of these tips with 15 continuing to use them at least some of the time. 81% of the participants reported continuing to use the tested protocols at least three times a week, with 43.8% reporting use of protocols every day. When asked if they had continued to use the App post-study, only four reported yes, with one of them using it for a month “to try other experiments,” two used it for a few days, and one for an additional week, suggesting the main and lasting effect was created during the Phase B of the study. There is also evidence, however, that people explored practices from the XB experiments after the study. For example, the top three most chosen experiments from Phase B were, from Table 2:

L1—increasing bright light exposure during the day (19), S2—sleeping no less than 7 h per night (17) and C1—not drinking caffeine within 6 h of going to sleep (12). Of the Phase C respondents, while 12 reported getting more than 7 h of sleep/night since the study (Figure 5), only seven followed that protocol (S2) during Phase B. More broadly, 75% of respondents agreed that they were sleeping better. Significantly, all but one of the Phase C respondents attributed these changes to their engagement with the XB SleepBetter study.

Qualitative Reflections on Experiment in a Box Experience

Two of key questions for us were: 1) Are these small doses of *how* to use these protocols to *feel better* in the XB sufficiently strong to have a perceived immediate benefit; and then 2) Are they sufficiently strongly felt to maintain these practices over time and without the app? The Phase C data suggests yes, at least among those who responded to our 7-months request (64% of our sample): XBs seem to support experience, effect, practice over time, and that once learned, practices are accessible enough not to be forgotten. Our post Phase B interviews and Phase C open questions, reported below, were designed to explore which attributes of the XB interaction were experienced as supporting these.

In an effort to be as unbiased as possible in the interviews, we asked about initial expectations of the app, what interested people in participating, and also asked about the experience with each section of the app: aspects that worked; what could be improved; length of an experiment; reasons for switching experiments or not; and what participants feel they learned about themselves during the study. The following uses ID-codes for participants.

Experience in a Box

The inspiration of our work has been to make a connection between the lived experience of *feeling better* as a result of testing a health protocol. After 7 months, 50% of our Phase C respondents said the most important aspect of what XB facilitated was the

TABLE 3 | Statistical computations between the beginning of Phase B and end-of-B on perceived state questions.

		times waking up	Night terrors	Fall asleep	Wake up	Fresh	Sad	Sleepy	Tired	Stressed	Irritable	Concentration	Coordination	Appetite
Beginning of Phase B	Mean	2.36	1.48	2.24	3.24	2.72	2.92	2.68	3.08	2.6	2.36	2.4	2.24	3.56
	SD	1.28	0.58	0.87	1.30	1.17	1.49	1.14	0.95	1.35	1.07	0.91	1.01	1.15
End of Phase B	Mean	1.84	1.12	2.08	1.84	2.84	2.08	1.68	1.8	1.64	1.48	2.08	2	2.64
	SD	1.06	0.33	1.41	1.10	0.98	1.52	0.90	0.95	1.07	0.91	1.22	1.15	1.31
p-value		0.085	0.004	0.537	<0.001	0.559	<0.001	<0.001	<0.001	0.011	0.004	0.235	0.313	0.001
t-value		1.797	3.166	0.641	4.287	-0.592	3.674	4.201	4.369	2.753	3.226	1.218	1.030	3.663
Mean difference		0.52	0.36	0.16	1.4	-0.12	0.84	1	1.28	0.96	0.88	0.32	0.24	0.92

The bolded columns are the aspects which showed significant difference between the beginning of Phase B and end-of-B.

opportunity to “try a sleep practice, rather than being told it’s something you can do.” When asked what aspects of the app let them learn about sleep, which they would not otherwise have learned, they pointed to their well-being results from 1) experiencing how even very small changes can make a difference; 2) how important sleep is to performance, and 3) what doesn’t make a difference: A9—“I also tried other experiments, but they didn’t make any big impact so I didn’t really stick to them.”

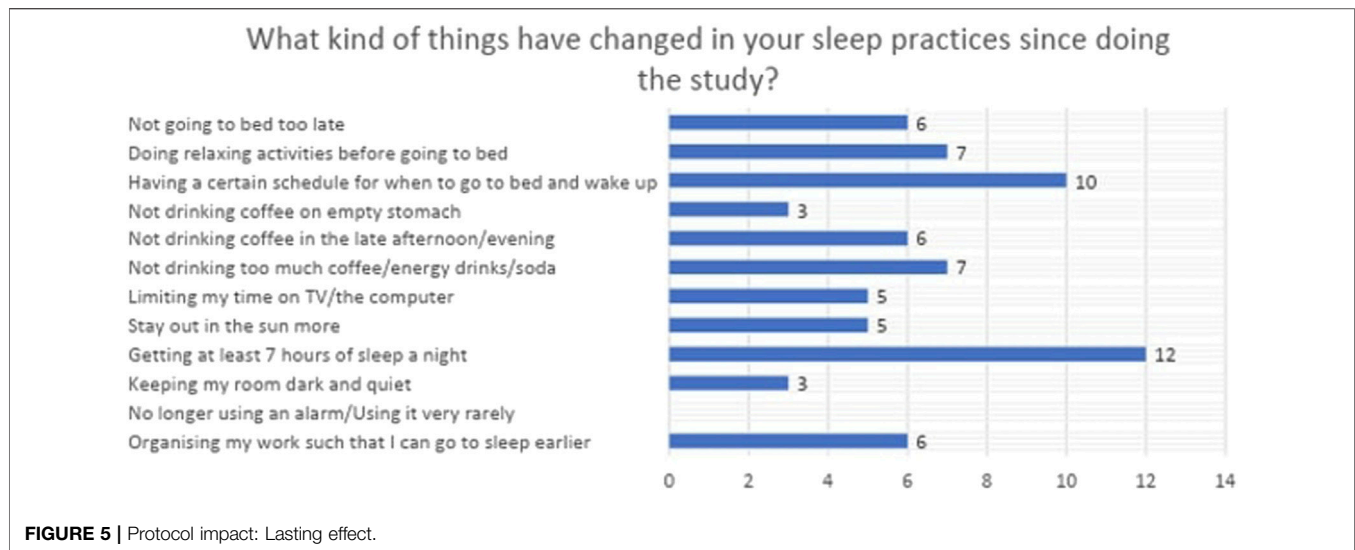
Pre-Fabricated Experiments to Build Experience Access

Overall, the opportunity to choose from a set of protocols to test was received positively. As A9 put it: “It was nice that I could try different experiments and see which is working better for me.” In other cases, participants mapped to experiments they thought were most related to their sleep issue—A13: “I chose the experiment based on what I thought my problem was.” A16 said, “I chose the ones that would be related to my everyday needs and activities.” One feature—the opportunity to select experiments in comparison to be “told what to do” was seen as particularly valuable—A4: “I think it is a good idea to let people choose instead of forcing something onto them because different things work for different people, letting them discover what works for them and then giving them the chance to stick with it if they found something good, or switch to something else if it didn’t work.” A2 framed this as empowering: “Because I was the one choosing the experiment made me really more motivated so it’s not like someone was telling me what to do.” All ten experiments could have a similar effect/benefit, but enabling participants to assess which might be easier for them to test or adopt was also a plus—A12: “making me choose which (experiment) I was more comfortable with would help me see which one affects my sleep the most, by going through each one at a time.” The length of time for the experiments also seemed appreciated. As A1 put it: “it gave participants enough time to get used to the experiment and feel a difference and be able to decide which one suits them better.” Implicit within these comments is also an appreciation that they are not being told what works but are testing effects for themselves.

Future Experiment Options

Some participants suggested that they had such a benefit from an experiment that they wanted to keep using it even while trying out another experiment in the following round. A12 stated that “after seeing that one experiment worked, I tried to keep it even though I switched to another one.” Indeed, after the first 5 days, 73.1% of the participants switched to another experiment. However, on the second experiment change, 69.6% remained with their current experiment for the last 5 days. Many of these said they would have liked to *stack experiments together*.

Some participants recommended that the app narrow down the set of experiments offered to make the selection even less of a choice, based on captured personal information. A13 suggests to “add a feature that would make the app choose the experiment for the user, based on their answers in the surveys and evolution.” Others, however, wanted an even wider set to find experiments even more in keeping with their daily lives. As A2 observes: “one thing to change about the app is the number of experiments available and the number of suggestions that the user has



because for me personally half of the experiments were not applicable so I had to choose between only two or three options.”

Connecting Practice With Experience

A goal in the XB design was to create mechanisms to assist, and potentially accelerate connecting the experience of a health practice into the value of adopting that practice. A critical component of building this experience was what we called *guided self-reflection*. We offered several mechanisms for this support: 1) the daily questionnaires, 2) the experiment-completion overview questionnaire that occurred on the fifth day of each experiment, 3) the notification reminders to complete the questionnaires and check the FAQ, and 4) the data visualisations.

We were surprised by the number of people who explicitly mentioned the once-a-day guided self-reflection practice as particularly helpful. Talking about the daily questionnaire, B5 and A9 reflected the general view: “the best part of the application was that at each end of the day you had one questionnaire to check your progress”; B5: “I guess the Questionnaire was the part I used the most” (also stated by A9). Several design factors seemed to contribute to the experience. The brevity of the questions was seen as a plus. As A17 said: “the questionnaire was not too long and it did not take too much time.” A2 appreciated the portability the app offered for the study: “you can actually take the experiment home with you so you didn’t have to go anywhere or talk to anyone.” Intriguingly, in Phase C, when asked what features of the XB people valued the most, only 12.5% reported that it was the questionnaires.

For others, the data visualization of the questionnaire responses over time was their main way to reflect on their experience. A17 reflects this general sentiment: “It helped me a lot because I was able to visualize my progress and see what I still have to work on.”

With respect to the Goal Diary, 14 people of the 25 used it. These entries would describe most often how participants felt about their progress, or something out of the ordinary affecting

their progress, like the succinct “Mosquito woke me up” of A4. Some of the users, like A1 have used the diary to record their actions (by adding how much they slept each day—“Slept 7h”). Others used it for their own recommendations based on their experience: A15 has logged in “Should go to sleep earlier.” and A10 has set their own goal “I will wake up early every day.”

The Calendar was also used to check engagement: A1 stated “the calendar was most helpful because it shows a recap of the day.” The ability to hover over a date to get more information about the experiment on that day proved useful as well: “I was able to check my status at any point in time” (A12).

Making Felt Connections

A key part of our work was to understand how *feeling* better affected perceptions of experience. When asked about their takeaways from the experience with the app, a sense of visceral experience was common. A1 said “I noticed that when my goal was reached I could concentrate more and feel less sleepy throughout the day.” Similarly, B5 said, “I learned that I could have a better sleep, and I can feel more rested and relaxed. And I saw that I can have a certain routine and respect it.”

There were also specific insights as well in terms of self-knowledge to inform future practice choices. A17 notes, “(the light experiment) helped me realize that my concentration really depends on the time I spend outside in the sunlight because it made me more focused and more positive.” Likewise, A10 learned about coffee timing’s effect on sleep, reporting “Even after finishing the experiment, I have given up my afternoon coffee, and kept that up.”

Design Suggestions

We also received a number of suggestions about how to improve the app, in particular to bring in more personal informatics, on demand. B5 suggested using a graph to display physical activity, heart rate or nutrition, whereas A4 requested to be able to track external factors affecting a protocol. Several participants wanted to track their hours slept more specifically throughout the experiment; others also

wanted to integrate other health practices like water or exercise.

In the Phase C responses, only three of the 16 respondents said they would have appreciated a longer intervention to have support to translate a practice into a habit. This request explicitly ties into our research question: do these experiences help build bridges/motivation toward adopting these healthier practices? The answer seems to be a qualified yes: for the majority of those who responded, they report ongoing practice without additional digital support. For these few others, they want to sustain the practice with more assistance. There is a role potentially here for building synergies from XB toward apps that then specifically support a tested and desirable practice.

Study Two Summary

It's clear from a comparison between Phases A and B, that, in this participant sample, the experience of using an XB was effective in having a perceived benefit on participants' sleep. In Phase C, it also seems clear that the period of the intervention—15 days—was sufficient to establish knowledge, skills, and practice that continued to be used, without further aid of technology, for at least half a year post-intervention.

Study Three: Offering Knowledge and Generic Practice vs. Knowledge and Practice Plus Self-Study Skills

A key question stemming from the perceived utility of XBs in the previous studies is it the features of the XB creating this effect, or is it simply giving people what may be new information and encouragement to try out these health skills? The XB Model provides a structure to support self-regulatory skills of self-study plus knowledge and practice encapsulated in generic health heuristics. Hence, we had the following hypothesis:

The XB Model would be more effective at promoting improved sleep quality compared to providing knowledge, health behavioral skills (arguably current standard of care) while not being taught the self-study skills encapsulated in an XB.

Specifically, the common approach within health communication is to offer knowledge based on prior science and evidence-based recommendations and then offer concrete suggested activities and skills that one could enact as a regular practice. For example: Caffeine can interrupt sleep (Knowledge). Based on this, avoid caffeine 6 h prior to bedtime (Practice). The key distinction between this and the XB model is two-fold: 1) how health heuristics are presented; and 2) helping people build the skills of self-study. The XB does not differ much in terms of the knowledge it is grounded in. It does make adjustments though in how to offer a health heuristic: specifically, we see it as 1) critical to frame health heuristics that inspire curiosity, 2) actionable, 3) can result in meaningful effects relatively quickly, and 4) the health heuristic is offered in a way that inspires a person to not enact it as a habit (which is often the implied target for generic messaging), but instead to experiment with the heuristic. Further, in terms of skills, the XB model provides people a structure to translate a generic health heuristic into an actionable personal health heuristic.

Procedure

To explore this feature, we designed a between-subjects study in which participants were randomized into either the control condition, in line with standard of care, or into our XB model. The study ran again for 1 week set up, with a baseline questionnaire about sleep quality, a 15-days execution with the invitation to run one to three experiments per 5 days/experiment. We deployed two versions of our SleepBetterXB app: the same app used in study two (now both Android and iOS) vs. an app that provided them knowledge about sleep and offered them basic sleep heuristics, with a suggestion to try it for 5 days. During the 5 days, the person would be provided with the health heuristic and be informed on how many days were left trying it out. After 5 days, people were suggested to either keep doing the same health heuristic or try a new one. In both groups, this was a 15-days study, with three experiments. Like study two, participants were recruited *via* social media, but went outside the university zone, deliberately also recruiting for gender balance across conditions.

Findings

In the XB condition, there were 16 participants: eight men and eight women; in the Control condition there were 15: 7 men, seven women and one Prefer Not to Say. No one was paid to participate or compensated for their participation. Our two-target outcomes were completion and sleep quality. Completion was defined as engagement in all aspects of the experience, including using the app daily and completing the sleep quality assessments, particularly at baseline and after the 15-days study. Based on this definition, four out of 15 (26.6%) participants completed the control condition compared to 15 out of 16 (93.4%) in the XB group. Central to this, most participants in the control stopped engaging with the app within the first 5 days.

Given the significant lack of completion in the control group, we did not run any statistical analysis related to sleep quality in that group. The fact that the control abandonment was so high and happened within the first 5 days of the study launch indicated to us that there is a perceived difference in experience, thus establishing a strong justification to not use statistics as the between-group differences are clear and strong without inferential statistics. In line with study two, a paired *t*-test of pre/post intervention with the XB groups showed a significant effect ($PreMean = 2.9 \pm 0.63$, $PostMean = 2.42 \pm 0.60$, $t = 2.186$ with $df = 14$, and $p = 0.046$).

Overall, these results indicate a strong difference in completion between the control and the XB apps. Further, results replicate the study two findings, suggesting that the XB app—in particular its features supporting engagement and reflection—facilitated improvements in sleep quality.

SUMMARY OF FINDINGS ACROSS THREE STUDIES—A CASE FOR CONSILIENCE

As a reminder, the goals of our three studies were to advance the key goals of phase I and phase IIa of the ORBIT model. Specifically, this means that our goal was to define and refine the XB approach and then to test if we were getting a meaningful signal. If a clear

signal is observed, this would be robust initial evidence that would justify more rigorous clinical trials in later phase II and phase III trials within ORBIT. Overall, the studies enabled us to iteratively define and refine the XB approach, to enable it to be more formalized and repeatable across three studies, with relatively minimal changes between study two and three. Further, our series of proof-of-concept studies suggest that, indeed, there is a signal of benefit for our approach and illustrate a general pattern of benefit of the XB approach across three different groups and, between study one vs. two and three, different XBs. Further, results from study three, particularly with the strong drop-off, which is consistent with standard use of most mobile health apps [e.g. (Bot et al., 2016)]. Taken together, these results highlight the value of supporting people explore health practices, and to enable people to test these practices for personal efficacy in terms of the simple heuristic: *how do you feel: better or worse or the same today than yesterday?* And then to have experience of different skills they might bring to bear to feel better.

In particular, the concept of XB as an *experiment*, a time-limited exploration of health heuristics is effective at making it safe to explore these practices *via its “try it first before committing”* approach (Study One). We see also see that XB’s translation into digital interventions provide a sufficient foundation for the practices tested to endure initial exposure to the app, and then, as we had hoped would be the case, to be able to continue these practices without the app (Study Two). We also have strong indication that the XB attributes that foster active practice and deliberate reflection are key to people engaging, exploring and building these personal heuristics (Study Three). Taken together, we conclude that there is sufficient evidence to warrant further exploration and, subsequently, more rigorous experimental (in the sense of the word used in health sciences) evaluation of the approach.

THE PROPOSED EXPERIMENT IN A BOX FRAMEWORK

From the results of this work, we propose seven components from the XB design components of generalized XB framework as the foundation for future work:

1. *Experiment as Conceptual Framing.* The framing of the interaction as an *experiment* has multiple traits that show up as contributing to the overall utility of the approach. Aligned with self-determination theory (Stevens et al., 2016), an experiment supports autonomy, competency and relatedness. It supports autonomy by enabling the person to choose and run their own experiment, and to respect their own determination about it. It supports competency by first asserting that their determination about the intervention’s utility is valid, and it helps them build new KSP about healthful choices that work for them. The experiment as a frame also emphasizes the approach being tested, not the person. Participants sharing their experiences/data from the experiments, as we did in Study One, also helps build relatedness. This approach we are seeing is allowing us to frame difficult questions related to sense of healthy selves, too. We are collaborating with an area college to

engage students’ interest in their own personal sleep practice experiments’ data for multiple curriculum purposes, but grounded in a provocation these students asked to explore: “am I normal?”—what does that mean, and how explore it. Thus, there is motivation to explore statistical to social science curricula 1) to makes sense of one’s own findings relative to their cohorts’ and other national groups 2) to problematize with these live examples how statistical “norms” intersect with and can conflict with associated cultural practice norms 3) to explore how these results and concepts can be constructed as healthy or not 4) to explore complexities of policy making relative to the uses of such qualitative and quantitative data.

2. *Multiple, Pre-Fabricated Heuristic Experiments:* There are two aspects to this design factor. One, we offer multiple paths to the same experience: there is, for example, more than one way to get a better night’s sleep. That multiplicity in itself is a signal that there is a lot of choice around how to achieve a health win. Indeed, as we saw, a few of our participants wanted that choice space, not to go away, but in fact to be narrowed. That request may be an opportunity for machine learning eventually to pick up more personal parameters to reduce the set of candidates. Second, the experiment here is being used to enable what is frequently a new experience. We are using that *experience as a bridge into new practices*. To this end, we use already-known science to inform health heuristics, rather than creating new science. This focus is in contrast to most current self-experimentation work that seems to be used to create new answers to solve a perceived problem, whether via large scale N-of-1 studies (Pandey et al., 2017), or within a single-case study (Karkar et al., 2015). This is hugely valuable and important work. In our case, our focus is simply different, though we think strongly complementary. Though there are opportunities for these informal XB experiments to contribute to science, that is not our innovation focus here. We want to help people find, connect-with and test *existing* solutions-as-heuristics to build resilient health practices. We translate established protocols into experiments so that a person can sample *for themselves* what may be an entirely novel experience, and to assess it in terms of simplicity—but perhaps profoundly—*how they feel*—as opposed to any formal effects measured in blood chemistry. In this way, we are helping them create not only embodied signal sensitivity but providing and emphasizing that they have options on what to do about the signal.

3. *Control and Choice* The simple opportunity to choose for oneself from a range of strategies to feel better was roundly applauded. Having the opportunity to try several or to stick with one method over 15 days was another choice factor that was a win. The approach also somewhat mirrors/reverses SleepCoacher (Daskalova et al., 2016), where external sources are vetting sleep data and suggesting protocol refinements. These related strategies suggest there may be times and conditions under which one approach is a better starting point than another. We suspect that there is value, as the MAHI work showed (Mamykina et al., 2008), in enabling a sense of empowerment in exploring health paths.

4. *Guided Self-Reflection:* Self-reflection is a key aspect of health self-study (Li et al., 2011). In our case, we built in what we foreground as *guided* self-reflection to help steer participants to focus on key parts of their experience with a given protocol

relative to its perceived effect on their overall personal state and their desired state. We saw quite strongly that offering multiple types of guided reflections on the practice was effective. Participants appreciated the daily notifications and pauses to reflect on 1) how the protocol they were testing was being experienced, and 2) for a number of our participants, on what this seemed to tell them, beyond the experiment, about their lifestyles.

5. *Time*: Experiments run for a set (and we think, critically, short) period to test an effect of a specific protocol. As we saw with our participants, it is not only far easier for someone to commit to do some protocol for a minimal period like 5 days than 12 weeks; it is also possible to feel the effect of an intervention and make a judgment about it in that time. Indeed, a key design marker for experiments is: what is the evidence that an experience can be felt within this window? If there is such support, it is a good candidate for an XB. A related aspect around self-tracking may be that: more people may be willing to use something when the tracking is seen to be for only a short time, toward a very clear purpose like assessment of a practice. This could have an added benefit of creating new data resources for further research benefit.

6. *Try It: All Data Are Good Data*: The opportunity to test out practices was the most valued part of the XB. This “*try before you buy*” is very distinct from the dominant prescriptive approach in health to *do* a specific thing—usually for the rest of one’s life, if time is specified at all. This exploratory approach switches the dynamic to say that the person is in control of this process, not the other way around: the person can accept or reject a protocol based on a very simple heuristic: “Do you *feel* better or worse after doing it?” The XB approach signaled strongly also that the person was *assessing* these practices: if they were not effective, it was not the person’s fault. That lack of blame also seems empowering in these designs. This is echoed by A12 who in the Phase C survey said “*It (doing the XB study) made me curious of more sleeping practices that I can try. Since I bought a google assistant, I have also been listening to the sound of rain while sleeping. This has significantly improved my sleep and I don’t think I would have been willing to explore this if I hadn’t been given the chance to try new sleep practices.*” We plan to explore designs to more deliberately support developing just this kind of confidence.

7. *You Are Your Own Sensor*. A guiding principle for the XB is to create personal health heuristics that can outlast a technology. To that end, the heuristics we are building are ones that will support health without requiring technology beyond one’s senses. When we ask “*How do you feel?*” as the main question, we aim to connect the sensory-oriented heuristic with the visceral practice. With a range of health heuristics, they will also have the start of building fundamentals, as we have seen from our participants to debug their experiences when they feel better or worse.

FUTURE WORK

The work presented here showed XBs were successful in helping the people who used them build resilient health heuristics across three exploratory studies. With this, future work is needed to do more rigorous evaluation of the work. In addition, greater study of the

proposed XB Framework elements would be valuable: are all seven of the proposed XB framework components equally necessary to create this effect as a gestalt? For example, while SDT requires “relatedness” as part of best support for engagement, Study Two did not have that feature, and yet participants engaged and built lasting skills. Future work should examine how study participation itself may have played a role in enacting a sense of relatedness, as implied by self-determination theory (Stevens et al., 2016). Future work may tease apart these attributes to better understand when and where any of these attributes may be more critical. Various participants have requested being able to “stack” experiments. We see this as valuable, as participants could blend the in five with their own variables. For instance, a participant may wish to put movement and sleep together with consideration of light and air quality: does movement indoors vs. outdoors affect sleep quality or daytime alertness? We are also keen to support group use for XBs to explore how collective health practices may inform health and wellbeing cultures in groups, and where the experimental outputs may be used as evidence/support to underpin organizational decision making as co-design rather than performance monitoring. Fundamentally, we are most keen to see how others might use the XB framework to create new experiment boxes as we anticipate that the framework can generalize to other life-skill learning, beyond health, as well.

CONCLUSION

Our three studies produced both a clear definition of the XB framework plus also, across three proof-of-concept studies, a clear signal of effect. Together, this work establishes the XB framework and provides justification for further research and investment in it. In particular, we propose a newly proposed continuum across two dimensions: *outsourcing to insourcing* and *habit to heuristic* to guide and interpret development and testing of interactive health technologies. The XB lets us ask, as a design question, where the ownership of a health practice is being developed: trust in the device/prescription in outsourcing, or knowledge, skills, and practice-building in the individual? We suggest this framing may help create new kinds of health interactions in particular that may be attractive to the known groups of users for whom health tracking is a non-starter or has been abandoned, but for whom engagement in health practices would be a benefit. This work also proposed health heuristics as a complement to the more familiar concept *health habits*. The key to this distinction is that a personal health heuristic enables a person to adapt their practice to different contexts. This contextual resilience is essential as it aligns well with the evolving definition of health (Nobile, 2014) as a process of adapting, of dynamic resilience, rather than state.

Key contributions of this work include the XB framework to instantiate health “Experiments in a Box” to help individuals explore, evaluate and develop personal health heuristics. We offer several formative studies that demonstrate how the XB model can be instantiated both broadly across health topics and more deeply within one health topic. The XB is framed as an “experiment,” as the word is used colloquially, to foreground the opportunity to evaluate the utility of a heuristic and that it may or may not work

for an individual as a way to help normalize this evaluation as part of building a health practice. We frame this experimental engagement as translating a general health heuristic to a personal health heuristic. We show that with even short engagement with the XB process, participants develop health practices from the interventions that are still in use long after the intervention is finished. Overall, the work offers a new scaffolding by which to frame time-limited interactive technology interventions to support building the knowledge skills and practice that can thrive in that person, significantly both post-interventions, and independent of that technology.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article can be made available by the authors, upon request, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ERGO Ethics and Research Governance Office (43589.A1), University of Southampton. The patients/participants provided their written informed consent to participate in this study. ERGO Ethics and Research Governance Office, University of Southampton (43589.A1).

REFERENCES

- Agapie, E., Colusso, L., Munson, S. A., and Hsieh, G. (2016). "PlanSourcing," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. New York, NY, USA: ACM, 119–133. doi:10.1145/2818048.2819943
- Ajana, B. (2017). Digital Health and the Biopolitics of the Quantified Self. *Digital Health* 3, 205520761668950. doi:10.1177/2055207616689509
- Allen, Hunter., Coggan, Andrew. R., and McGregor, Stephen. (2019). *Training and Racing with a Power Meter*. Boulder, CO, USA: VeloPress.
- Ameen, Mohamed. S., Cheung, Lok., Hauser, Theresa., Hahn, Michael. A., and Schabus, Manuel. (2019). About the Accuracy and Problems of Consumer Devices in the Assessment of Sleep. *Sensors* 19 (19), 4160. doi:10.3390/s19194160
- Barcena, Mario. Ballano., Wueest, Candid., and Lau, Hon. How Safe Is Your Quantified Self? *SECURITY RESPONSE* 38.
- Baron, K. G., Abbott, S., Jao, N., Manalo, N., and Mullen, R. (2017). Orthosomnia: Are Some Patients Taking the Quantified Self Too Far? *J. Clin. Sleep Med.* 13 (2), 351–354. doi:10.5664/jcsm.6472
- Barry, V. W., Baruth, M., Beets, M. W., Durstine, J. L., Liu, J., and Blair, S. N. (2014). Fitness vs. Fatness on All-Cause Mortality: A Meta-Analysis. *Prog. Cardiovasc. Dis.* 56 (4), 382–390. doi:10.1016/j.pcad.2013.09.002
- Bauer, J., Consolvo, S., Benjamin, G., Schooler, J., Wu, E., Watson, N. F., et al. (2012). "ShutEye," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. New York, NY, USA: ACM, 1401–1410. doi:10.1145/2207676.2208600
- Bechara, A., Damasio, H., and DamasioDamasio, A. R. (2003). Role of the Amygdala in Decision-Making. *Ann. N Y Acad. Sci.* 985, 356–369. doi:10.1111/j.1749-6632.2003.tb07094.x
- Bentley, F., Tollmar, K., Stephenson, P., Levy, L., Jones, B., Robertson, S., et al. (2013). Health Mashups. *ACM Trans. Comput.-Hum. Interact.* 20 (5), 1–27. doi:10.1145/2503823

AUTHOR CONTRIBUTIONS

MS developed the XB approach and study one; GM co-designed and lead study two; EH co-designed study three and contributed to analysis and organization of results.

FUNDING

UKRI EPSRC Health Resilience Interactive Technology, ReFresh and GetAMoveOn (No. EP/T007656/1, EP/K021907/1, and EP/N027299/1) are various government funded projects, the work from which has enabled the exploration of the ideas in this paper.

ACKNOWLEDGMENTS

We acknowledge support from the UKRI EPSRC, in particular, the Health Resilience Interactive Technology Established Career Fellowship, the ReFresh Project and the GetAMoveOn Health Network+ (EP/T007656/1, EP/K021907/1, EP/N027299/1). We also acknowledge support from the office of the Pro Vice Chancellor, Interdisciplinary Research, University of Southampton, United Kingdom. Finally, we are grateful to the reviewers whose excellent feedback only enhanced this work.

- Destro, B., Paolo Anagnostou, G., Capocasa, M., Bencivelli, S., Cerron, A., Contreras, J., Enke, N., et al. (2014). *Perspectives on Open Science and Scientific Data Sharing: An Interdisciplinary Workshop* (Rochester, NY: Social Science Research Network.). SSRN Scholarly Paper ID 2456712. <https://papers.ssrn.com/abstract=2456712>
- Blandford, A. (2019). HCI for Health and Wellbeing: Challenges and Opportunities. *Int. J. Human-Computer Stud.* 131, 41–51. doi:10.1016/j.ijhcs.2019.06.007
- Bot, B. M., Suver, C., Neto, E. C., Kellen, M., Klein, A., Bare, C., et al. (2016). The MPower Study, Parkinson Disease Mobile Data Collected Using ResearchKit. *Sci. Data* 3 (1), 1–9. doi:10.1038/sdata.2016.11
- Brooks, Richard (2020). *The Failure of Test and Trace Shows the Folly of Handing Huge Contracts to Private Giants* | Richard Brooks'. London, United Kingdom: The Guardian. <http://www.theguardian.com/commentisfree/2020/oct/13/failure-test-trace-folly-huge-contracts-private-giants-uk>.
- Buman, M. P., Giacobbi, P. R., Dzierzewski, J. M., Morgan, A. A., McCrae, C. S., Roberts, B. L., et al. (2011). Peer Volunteers Improve Long-Term Maintenance of Physical Activity with Older Adults: A Randomized Controlled Trial. *J. Phys. Activity Health* 8 (s2), S257–S266. doi:10.1123/jpah.8.s2.s257
- Chan, R., Shum, D., Touloupoulou, T., and Chen, E. (2008). Assessment of Executive Functions: Review of Instruments and Identification of Critical Issues. *Arch. Clin. Neuropsychol.* 23 (2), 201–216. doi:10.1016/j.acn.2007.08.010
- Chevance, G., Perski, O., and Hekler, E. B. (2020). Innovative Methods for Observing and Changing Complex Health Behaviors: Four Propositions. *Translational Behav. Med.* 11 (2), 676–685. doi:10.1093/tbm/ibaa026
- Choe, E. K., Lee, B., Kay, M., Pratt, W., and Kientz, J. A. (2015). "SleepTight," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. New York, NY, USA: ACM, 121–132. doi:10.1145/2750858.2804266
- Choe, E. K., Lee, B., and schraefel, m. c. (2015a). Characterizing Visualization Insights from Quantified Selfers' Personal Data Presentations. *IEEE Comput. Graph. Appl.* 35 (4), 28–37. doi:10.1109/MCG.2015.51

- Churchill, E. F., and schraefel, m. c. (2015). MHealth + Proactive Well-Being = Wellth Creation. *Interactions* 22 (1), 60–63. doi:10.1145/2690853
- Clawson, J., Pater, J. A., Miller, A. D., Mynatt, E. D., and Mamykina, L. (2015). “No Longer Wearing,” in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. New York, NY, USA: ACM, 647–658. doi:10.1145/2750858.2807554
- Consolvo, S., Klasnja, P., McDonald, D. W., Avrahami, D., Froehlich, J., LeGrand, L., et al. (2008). “Flowers or a Robot Army?,” in *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp '08)*. New York, NY, USA: ACM, 54–63. doi:10.1145/1409635.1409644
- Czajkowski, S. M., Powell, L. H. Nancy. Adler, N., Naar-King, S., Reynolds, K. D., Hunter, C. M., et al. (2015). From Ideas to Efficacy: The ORBIT Model for Developing Behavioral Treatments for Chronic Diseases. *Health Psychol.* 34 (10), 971–982. doi:10.1037/hea0000161
- Daskalova, N., Desingh, K., Papoutsaki, A., Schulze, D., Sha, H., and Huang, J. (2017). Lessons Learned from Two Cohorts of Personal Informatics Self-Experiments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1 (3), 1–22. doi:10.1145/3130911
- Daskalova, N., Metaxa-Kakavouli, D., Tran, A., Nugent, N., Boergers, J., McGeary, J., et al. (2016). “SleepCoacher,” in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. New York, NY, USA: ACM, 347–358. doi:10.1145/2984511.2984534
- Dinis, J., and Bragança, M. (2018). Quality of Sleep and Depression in College Students: A Systematic Review. *Sleep Sci.* 11 (4), 290–301. doi:10.5935/1984-0063.20180045
- Duan, N., Kravitz, R. L., and Schmid, C. H. (2013). Single-Patient (N-of-1) Trials: A Pragmatic Clinical Decision Methodology for Patient-Centered Comparative Effectiveness Research. *J. Clin. Epidemiol.* 66 (8), S21–S28. doi:10.1016/j.jclinepi.2013.04.006
- Dulaud, P., Di Loreto, I., and Mottet, D. (2020). Self-Quantification Systems to Support Physical Activity: From Theory to Implementation Principles. *Ijerp* 17 (24), 9350. doi:10.3390/ijerp17249350
- Epstein, D. A., Caraway, M., Johnson, C., Ping, A., Fogarty, J., and Munson, S. A. (2016). “Beyond Abandonment to Next Steps,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. New York, NY, USA: ACM, 1109–1113. doi:10.1145/2858036.2858045
- Erbas, Y., Ceulemans, E., Kalokerinos, E. K., Houben, M., Koval, P., Pe, M. L., et al. (2018). Why I Don't Always Know what I'm Feeling: The Role of Stress in Within-Person Fluctuations in Emotion Differentiation. *J. Personal. Soc. Psychol.* 115 (2), 179–191. doi:10.1037/pspa0000126
- Espie, C. A., Kyle, S. D., Hames, P., Gardani, M., Fleming, L., and Cape, J. (2014). The Sleep Condition Indicator: a Clinical Screening Tool to Evaluate Insomnia Disorder. *BMJ Open* 4 (3), e004183. doi:10.1136/bmjopen-2013-004183
- Feldman, R. (2020). What Is Resilience: An Affiliative Neuroscience Approach. *World Psychiatry* 19 (2), 132–150. doi:10.1002/wps.20729
- Ferriss, Timothy. (2009). *The 4-Hour Workweek: Escape 9-5, Live Anywhere, and Join the New Rich. Expanded, Updated Ed. Edition*. New York: Harmony.
- Freedland, Kenneth. E. (2020a). Pilot Trials in Health-Related Behavioral Intervention Research: Problems, Solutions, and Recommendations. *Health Psychol.* 39 (10), 851–862. doi:10.1037/hea0000946
- Freedland, Kenneth. E. (2020b). Purpose-Guided Trial Design in Health-Related Behavioral Intervention Research. *Health Psychol.* 39 (6), 539–548. doi:10.1037/hea0000867
- Griffith, E. (2020). ‘Peloton: The Latest Virus Panic Buy’. *Chicagotribune.Com*. New York, NY, USA. 7 May 2020. <https://www.chicagotribune.com/coronavirus/sns-nyt-peloton-boom-workout-exercise-equipment-coronavirus-20200509-vu47jlv2fdkxapxtt6zx7lqe-story.html>.
- Haidt, Jonathan. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. New York, NY, USA: Vintage.
- Haraway, Donna. Jeanne. (1991). *Simians, Cyborgs, and Women: The Reinvention of Nature*. New York: Routledge. <http://archive.org/details/simianscyborgsw0000hara>.
- Hashemzadeh, M., Rahimi, A., Zare-Farashbandi, F., Alavi-Naeini, A. M., and Daei, A. (2019). Transtheoretical Model of Health Behavioral Change: A Systematic Review. *Iran J. Nurs. Midwifery Res.* 24 (2), 83–90. doi:10.4103/ijnmr.IJNMR_94_17
- Heffernan, Margaret. (2013). *What Happened after the Foxconn Suicides*. New York, NY: CBSNews. CBS Interactive.
- Hekler, E. B., Buman, M. P., Rivera, Poothakandiyil. N., Dzierzewski, Joseph. M., Morgan, Adrienne. Aiken., Rivera, D. E., et al. (2013). Exploring Behavioral Markers of Long-Term Physical Activity Maintenance. *Health Educ. Behav.* 40 (1_Suppl. 1), 51S–62S. doi:10.1177/1090198113496787
- Hekler, E. B., Rivera, D. E., Martin, Cesar. A., Phatak, Sayali. S., Freigoun, Mohammad. T., Korinek, Elizabeth., et al. (2018). Tutorial for Using Control Systems Engineering to Optimize Adaptive Mobile Health Interventions. *J. Med. Internet Res.* 20 (6), e214. doi:10.2196/jmir.8622
- Hekler, Eric. B., Klasnja, Predrag., Chevance, Guillaume., Golaszewski, Natalie. M., Lewis, Dana., and Sim, Ida. (2019). Why We Need a Small Data Paradigm. *BMC Med.* 17 (1), 1–9. doi:10.1186/s12916-019-1366-x
- Hekler, Eric. B. (2013). “Winslow Burleson, and Jisoo LeeA DIY Self-Experimentation Toolkit for Behavior Change,” in *ACM Conference on Human Factors in Computing Systems*. doi:10.1145/2470654.2466452
- Hekler, E., Tiro, J. A., Hunter, C. M., and Nebeker, C. (2020). Precision Health: The Role of the Social and Behavioral Sciences in Advancing the Vision. *Ann. Behav. Med.* 54 (11), 805–826. doi:10.1093/abm/kaa018
- Heyen, N. B. (2020). From Self-Tracking to Self-Expertise: The Production of Self-Related Knowledge by Doing Personal Science. *Public Underst. Sci.* 29 (2), 124–138. doi:10.1177/0963662519888757
- Ibáñez, V., Silva, J., and Cauli, O. (2018). A Survey on Sleep Questionnaires and Diaries. *Sleep Med.* 42, 90–96. doi:10.1016/j.sleep.2017.08.026
- Ji, S., Cherry, C. R., Han, L. D., and Jordan, D. A. (2014). Electric Bike Sharing: Simulation of User Demand and System Availability. *J. Clean. Prod.* 85, 250–257. doi:10.1016/j.jclepro.2013.09.024
- Karkar, R., Fogarty, J., Kientz, J. A., Munson, S. A., Vilardaga, R., and Zia, J. (2015). “Opportunities and Challenges for Self-Experimentation in Self-Tracking,” in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers - UbiComp '15* (Osaka, Japan: ACM Press), 991–996. doi:10.1145/2800835.2800949
- Karkar, R., Schroeder, J., Epstein, D. A., Pina, L. R., Scofield, J., Fogarty, J., et al. (2017). TummyTrials. *CHI Conf.* 2017, 6850–6863. doi:10.1145/3025453.3025480
- Karkar, R., Zia, J., Vilardaga, R., Mishra, S. R., Fogarty, J., Munson, S. A., et al. (2016). A Framework for Self-Experimentation in Personalized Health. *J. Am. Med. Inform. Assoc. JAMIA* 23 (3), 440–448. doi:10.1093/jamia/ocv150
- Kay, M., Morris, D., schraefel, m., and Kientz, J. A. (2013). “There's No Such Thing as Gaining a Pound,” in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. New York, NY, USA: ACM, 401–410. doi:10.1145/2493432.2493456
- Kearney, Lauren. (2018). *20 Tactics Bear Grylls Uses to Survive the Wilderness (That Actually Work)*. Quebec, Canada: TheTravel. 17 August 2018. <https://www.thetravel.com/20-tactics-bear-grylls-uses-to-survive-the-wilderness-that-actually-work/>
- Kleckner, Ian. R., Zhang, Jiahe., Touroutoglou, Alexandra., Chanes, Lorena., Xia, Chenjie., Simmons, W. Kyle., et al. (2017). Evidence for a Large-Scale Brain System Supporting Allostasis and Interoception in Humans. *Nat. Hum. Behav.* 1 (5), 1–14. doi:10.1038/s41562-017-0069
- Kravitz, Richard. L., Aguilera, Adrian., Chen, Elaine. J., Choi, Yong. K., Hekler, Eric., Karr, Chris., et al. (2020). Feasibility, Acceptability, and Influence of MHealth-Supported N-Of-1 Trials for Enhanced Cognitive and Emotional Well-Being in US Volunteers. *Front. Public Health* 8, 260. doi:10.3389/fpubh.2020.00260
- Kwasnicka, D., Dombrowski, S. U., White, M., and Sniehotta, F. (2016). Theoretical Explanations for Maintenance of Behaviour Change: A Systematic Review of Behaviour Theories. *Health Psychol. Rev.* 10 (3), 277–296. doi:10.1080/17437199.2016.1151372
- Lee, J., Walker, E., Burleson, W., Kay, M., Buman, M., Hekler, E. B., et al. (2017). “Self-Experimentation for Behavior Change: Design and Formative Evaluation of Two Approaches,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA: Association for Computing Machinery), 6837–6849. doi:10.1145/3025453.3026038
- Lerman, Rachel. (2020). *Feeling Stressed? Meditation Apps See Surge in Group Relaxation*. Washington, DC: Washington Post. <https://www.washingtonpost.com/technology/2020/04/21/meditation-up-during-coronavirus/>.

- Li, I., Dey, A. K., and Forlizzi, J. (2011). "Understanding My Data, Myself," in *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp '11)*. New York, NY, USA: ACM, 405–414. doi:10.1145/2030112.2030166
- Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., and Barrett, L. F. (2012). The Brain Basis of Emotion: A Meta-Analytic Review. *Behav. Brain Sci.* 35 (3), 121–143. doi:10.1017/s0140525x11000446
- Lockton, D., Zea-Wolfson, T., Chou, J., Song, Y., Ryan, E., and Walsh, C. (2020). "Sleep Ecologies," in *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (New York, NY, USA: Association for Computing Machinery), 1579–1591. doi:10.1145/3357236.3395482
- Loke, L., and Schiphorst, T. (2018). The Somatic Turn in Human-Computer Interaction. *Interactions* 25 (5), 54–5863. doi:10.1145/3236675
- Louzon, P., Jessica, L. A., Torres, X., Pyles, E., Ali, M., Du, Y., et al. (2020). Characterisation of ICU Sleep by a Commercially Available Activity Tracker and its Agreement with Patient-Perceived Sleep Quality. *BMJ Open Res. Res.* 7 (1), e000572. doi:10.1136/bmjresp-2020-000572
- Luft, A. R., and Buitrago, M. M. (2005). Stages of Motor Skill Learning. *Mn* 32 (3), 205–216. doi:10.1385/mn:32:3:205
- Lunde, P., Nilsson, B. B., Bergland, A., Kværner, K. J., and Bye, A. (2018). The Effectiveness of Smartphone Apps for Lifestyle Improvement in Noncommunicable Diseases: Systematic Review and Meta-Analyses. *J. Med. Internet Res.* 20 (5), e162. doi:10.2196/jmir.9751
- Mamykina, L., Mynatt, E., Davidson, P., and Greenblatt, D. (2008). "Mahi," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. New York, NY, USA: ACM, 477–486. doi:10.1145/1357054.1357131
- Mander, B. A., Winer, J. R., and Walker, M. P. (2017). Sleep and Human Aging. *Neuron* 94 (1), 19–36. doi:10.1016/j.neuron.2017.02.004
- Manzar, M. D., BaHammam, A. S., SpenceHameed, D. W., Pandi-Perumal, S. R., Moscovitch, A., and Streiner, D. L. (2018). Dimensionality of the Pittsburgh Sleep Quality Index: A Systematic Review. *Health Qual. Life Outcomes* 16 (1), 89. doi:10.1186/s12955-018-0915-x
- Mary, A., Schreiner, S., and Peigneux, P. (2013). Accelerated Long-Term Forgetting in Aging and Intra-sleep Awakenings. *Front. Psychol.* 4, 750. doi:10.3389/fpsyg.2013.00750
- McKay, F. H., Cheng, C., Wright, A., Shill, J., Stephens, H., and Uccellini, M. (2018). Evaluating Mobile Phone Applications for Health Behaviour Change: A Systematic Review. *J. Telemed. Telecare* 24 (1), 22–30. doi:10.1177/1357633X16673538
- Milojevich, H. M., and Lukowski, A. F. (2016). Sleep and Mental Health in Undergraduate Students with Generally Healthy Sleep Habits. *PLOS ONE* 11 (6), e0156372. doi:10.1371/journal.pone.0156372
- Murphy, J., Catmur, C., and Bird, G. (2019). Classifying Individual Differences in Interoception: Implications for the Measurement of Interoceptive Awareness. *Psychon. Bull. Rev.* 26 (5), 1467–1471. doi:10.3758/s13423-019-01632-7
- Nebeker, Camille., Weisberg, Bethany., Hekler, Eric., and Kurisu, Michael. (2020). Using Self-Study and Peer-To-Peer Support to Change "Sick" Care to "Health" Care: The Patient Perspective. *Front. Digit. Health*, doi:10.3389/fdgh.2020.00002
- Nielsen, J., and Molich, R. (1990). "Heuristic Evaluation of User Interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '90)*. New York, NY, USA: Association for Computing Machinery, 249–256. doi:10.1145/97243.97281
- Nobile, Marianna. (2014). THE WHO DEFINITION OF HEALTH: A CRITICAL READING. *Med. L.* 33 (2), 33–40.
- Ogilvy Consulting UK. (2014). 5 Ogilvy Behavioural Science Wellbeing Experiment. https://www.youtube.com/watch?v=UOS7ykGgzeA&ab_channel=OgilvyConsultingUK. doi:10.1093/obo/9780199830060-0106
- Olśzewska, Alicja. M., Gaca, Maciej., Herman, Aleksandra. M., Jednoróg, Katarzyna., and Marchewka, Artur. (2021). How Musical Training Shapes the Adult Brain: Predispositions and Neuroplasticity. *Front. Neurosci.* 15, 630829. doi:10.3389/fnins.2021.630829
- Oulasvirta, A., Rattenbury, T., Ma, L., and Raita, E. (2012). Habits Make Smartphone Use More Pervasive. *Pers Ubiquit Comput.* 16 (1), 105–114. doi:10.1007/s00779-011-0412-2
- Pandey, V., Amir, A., Debelius, J., Hyde, E. R., Kosciolk, T., Knight, R., et al. (2017). "Gut Instinct," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17* (Denver, Colorado, USA: ACM Press), 6825–6836. doi:10.1145/3025453.3025769
- Phatak, Sayali., Chen, Elaine., Stephen, Schueller, Kravitz, Richard., Sim, Ida., Schmid, Christopher., et al. (2018). "Design of a Large-Scale Self-Experimentation Tool for Scientific Self-Explorations," in *12th EAI International Conference on Pervasive Computing Technologies for Healthcare - Demos* (New York: Posters, Doctoral Colloquium). doi:10.4108/eai.20-4-2018.2276355
- Pierce, J., Schiano, D. J., and Paulos, E. (2010). "Home, Habits, and Energy," in *Proceedings Of the SIGCHI Conference On Human Factors In Computing Systems, 1985–94* (New York, NY, USA: Association for Computing Machinery). doi:10.1145/1753326.1753627
- Pinder, C., Vermeulen, J., Cowan, B. R., and Beale, R. (2018). Digital Behaviour Change Interventions to Break and Form Habits. *ACM Trans. Comput.-Hum. Interact.* 25 (3), 1–66. doi:10.1145/3196830
- Prince, J. D. (2014). The Quantified Self: Operationalizing the Quotidian. *J. Electron. Resour. Med. Libraries* 11 (2), 91–99. doi:10.1080/15424065.2014.909145
- Prochaska, J. O., and Velicer, W. F. (1997). The Transtheoretical Model of Health Behavior Change. *Am. J. Health Promot.* 12 (1), 38–48. doi:10.4278/0890-1171-12.1.38
- Rahman, T., Czerwinski, M., Gilad-Bachrach, R., and Johns, P. (2016). "Predicting "About-To-Eat" Moments for Just-In-Time Eating Intervention," in *Proceedings of the 6th International Conference on Digital Health Conference (DH '16)*. New York, NY, USA: Association for Computing Machinery, 141–150. doi:10.1145/2896338.2896359
- Riley, W. T., Rivera, D. E., Atienza, A. A., Nilsen, W., Allison, S. M., and Mermelstein, R. (2011). Health Behavior Models in the Age of Mobile Interventions: Are Our Theories up to the Task? *Behav. Med. Pract. Pol. Res.* 1 (1), 53–71. doi:10.1007/s13142-011-0021-7
- Sabrina, Siddiqui., and Sadie, Gurman. (2021). Biden Signs Executive Order to Phase Out Federal Use of Private Prisons. *Wall Street J.* 26., 2021 January 2021, sec. Politics. <https://www.wsj.com/articles/biden-to-sign-executive-orders-to-phase-out-federal-use-of-private-prisons-11611682272>
- Sanches, P., Janson, A., Karpashevich, P., Nadal, C., Qu, C., Roquet, Claudia. Daudén., et al. (2019). "HCI and Affective Health," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. New York, NY, USA: Association for Computing Machinery, 1–17. doi:10.1145/3290605.3300475
- Schlarb, Angelika., Bihlmaier, Isabel., Hautzinger, Martin., Gulewitsch, Marco. D., and Schwerdtle, Barbara. (2015). Nightmares and Associations with Sleep Quality and Self-Efficacy Among University Students. *J. Sleep Disord. Manag.* 11–5. <https://pub.uni-bielefeld.de/publication/2910766>.
- schraefel, m. c., Andres, Josh., Liu, Abby. Wanyu., Jones, Mike., Murname, Elizabeth., Tabor, Aaron., et al. (2021). *Body as Starting Point 4: Inbodied Interaction Design for Health Ownership*. New York, NY: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. <https://wellthlab.ac.uk/inbodied4>.
- schraefel, m. c. (2014). "Burn the Chair, We're Wired to Move: Exploring the Brain/Body Connexion for HCI Creativity & Cognition," in *Proceedings Of HCI Korea, 153–61. HCIK '15* (South Korea: Hanbit Media, Inc). <http://dl.acm.org/citation.cfm?id=2729485.2729509>.
- schraefel, m. c., and Hekler., E. (2020). Tuning. *Interactions* 27 (2), 48–53. doi:10.1145/3381897
- schraefel, m. c. (2019). "in5," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. New York, NY, USA: Association for Computing Machinery, 1–6. doi:10.1145/3290607.3312977
- schraefel, m. c. (2020). Inbodied Interaction: Introduction. *Interactions* 27 (2), 32–37. doi:10.1145/3380811
- schraefel, m. c., Tabor, A., and Andres., J. (2020). Toward Insourcing-Measurement in Inbodied Interaction Design. *Interactions* 27 (2), 56–60. doi:10.1145/3381340
- Sherwin, E., Rea, K., Dinan, T. G., and Cryan, J. F. (2016). A Gut (Microbiome) Feeling about the Brain. *Curr. Opin. Gastroenterol.* 32 (2), 96–102. doi:10.1097/MOG.0000000000000244
- Spence, S. (1995). Descartes' Error: Emotion, Reason and the Human Brain. *BMJ* 310 (6988), 1213. doi:10.1136/bmj.310.6988.1213

- Stawarz, K., Cox, A. L., and Blandford, A. (2015). "Beyond Self-Tracking and Reminders," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15. New York, NY, USA: ACM), 2653–2662. doi:10.1145/2702123.2702230
- Stevens, S. L., Wood, S., Koshiaris, C., Law, K., Glasziou, P., Stevens, R. J., et al. (2016). Blood Pressure Variability and Cardiovascular Disease: Systematic Review and Meta-Analysis. *Bmj* 354, i4098. doi:10.1136/bmj.i4098
- Sullivan, Alice., and Brown, Matt. (2013). *Overweight and Obesity in Mid-life: Evidence from the 1970 Birth Cohort Study at Age 42*. London, UK: Centre for Longitudinal Studies, Institute of Education, University of London. <http://discovery.ucl.ac.uk/10018085/>.
- The Sufferfest, 'The Sufferfest: Complete Training App for Cyclists and Triathletes'. n.d. Atlanta, GA: The Sufferfest. Accessed 29 January 2021. <https://thesufferfest.com/>.
- Thys, E., Sabbe, B., and De Hert, M. (2014). The Assessment of Creativity in Creativity/psychopathology Research - a Systematic Review. *Cogn. Neuropsychiatry* 19 (4), 359–377. doi:10.1080/13546805.2013.877384
- TrainingPeaks, 'TrainingPeaks | New Year. New Focus.' n.d. New York, NY: TrainingPeaks. Accessed 29 January 2021. <https://www.trainingpeaks.com/>.
- Troacă, Victor-Adrian., and Bodislav, Dumitru-Alexandru. (2012). Outsourcing. The Concept. *Theor. Appl. Econ.* 6 (6), 51.
- Tsakiris, M., and Critchley, H. (2016). Interoception beyond Homeostasis: Affect, Cognition and Mental Health. *Phil. Trans. R. Soc. B* 371, 20160002. doi:10.1098/rstb.2016.0002
- Tuominen, J., Peltola, K., Saaresranta, T., and Valli, K. (2019). Sleep Parameter Assessment Accuracy of a Consumer Home Sleep Monitoring Ballistocardiograph Beddit Sleep Tracker: A Validation Study. *J. Clin. Sleep Med.* 15 (3), 483–487. doi:10.5664/jcsm.7682
- Vandercammen, L., Hofmans, J., and Theuns, P. (2014). Relating Specific Emotions to Intrinsic Motivation: On the Moderating Role of Positive and Negative Emotion Differentiation. *PLoS One* 9 (12), e115396. doi:10.1371/journal.pone.0115396
- VerkhoshanskyY., Vitalievitch, and Cunningham Siff, Mel. (2009). *Supertraining 6th Ed. - Expanded Ed.* Rome, Italy: [Michigan: Verkhoshansky ; Distributed by Ultimate Athlete Concepts.
- West, J. H., Hall, P. C., Hanson, C. L., Barnes, M. D., Giraud-Carrier, C., and Barrett, J. (2012). There's an App for that: Content Analysis of Paid Health and Fitness Apps' There's an App for that: Content Analysis of Paid Health and Fitness Apps'. *J. Med. Internet Res.* 14 (3), e72. doi:10.2196/jmir.1977
- Wolf, Gary. Isaac., and De Groot, Martijn. (2021). A Conceptual Framework for Personal Science. *Frontiers in Computer Science* 2. <https://doi.org/10.3389/fcomp.2020.0002>.
- Wood, W., and Rünger, D. (2016). Psychology of Habit. *Annu. Rev. Psychol.* 67 (1), 289–314. doi:10.1146/annurev-psych-122414-033417
- Zucker, D. R., Schmid, C. H., McIntosh, M. W., D'Agostino, R. B., Selker, H. P., and Lau, J. (1997). Combining Single Patient (N-Of-1) Trials to Estimate Population Treatment Effects and to Evaluate Individual Patient Responses to Treatment. *J. Clin. Epidemiol.* 50 (4), 401–410. doi:10.1016/S0895-4356(96)00429-5

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.

Copyright © 2021 schraefel, Muresan and Hekler. 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.



What Does the Body Communicate With Postural Oscillations? A Clinical Investigation Hypothesis

Andrea Buscemi¹, Santi Scirè Campisi¹, Giulia Frazzetto¹, Jessica Petriliggieri¹, Simona Martino¹, Pierluca Ambramo¹, Alessandro Rapisarda¹, Nelson Mauro Maldonado², Donatella Di Corrado^{3*} and Marinella Coco⁴

¹Department of Research, Italian Center Studies of Osteopathy, Catania, Italy, ²Department of Neuroscience and Reproductive and Odontostomatological Sciences, University of Naples Federico II, Naples, Italy, ³Department of Human and Social Sciences, Kore University, Enna, Italy, ⁴Department of Biomedical and Biotechnological Sciences, University of Catania, Catania, Italy

OPEN ACCESS

Edited by:

Anna Esposito,
University of Campania Luigi
Vanvitelli, Italy

Reviewed by:

Tiziano A. Agostini,
University of Trieste, Italy
Matej Tusak,
University of Ljubljana, Slovenia

*Correspondence:

Donatella Di Corrado
donatella.dicorrado@unikore.it

Specialty section:

This article was submitted to
Human-Media
Interaction,
a section of the journal
Frontiers in Psychology

Received: 15 February 2021

Accepted: 13 May 2021

Published: 16 June 2021

Citation:

Buscemi A, Campisi SS, Frazzetto G, Petriliggieri J, Martino S, Ambramo P, Rapisarda A, Maldonado NM, Di Corrado D and Coco M (2021) What Does the Body Communicate With Postural Oscillations? A Clinical Investigation Hypothesis. *Front. Psychol.* 12:668192. doi: 10.3389/fpsyg.2021.668192

The evolution of the foot and the attainment of the bipedia represent a distinctive characteristic of the human species. The force of gravity is dissipated through the tibial astragalic joints, and the movement of the ankle is manifested on a sagittal plane. However, this is in contrast with other studies that analyze the straight station in bipodal support of the body. According to these studies, the oscillations of the body dissipated by the articulation of the ankle are greater on a frontal plane than on a sagittal plane. Probably, this can be deduced by analyzing the concept of “cone of economy (COE) and equilibrium;” a cone that has its base with the oscillations described by the 360° movement performed by the head and has its apex that supports polygon defined by the tibio-astragalic articulation. The purpose of this study was to evaluate a kind of communication between the oscillations of the COE and equilibrium and the main sphere of somatic dysfunction (structural, visceral, or cranial sacral), assessing the reliability of the “fascial compression test.” The implications of this connection have been considered, while grounding the hypothesis in the ability of the human body to maintain its center of mass (COM) with minimum energy expenditure and with minimum postural influence. At the same time, the fascial compression test provides a dominant direction of fascial compartments in restriction of mobility.

Keywords: body communication, postural oscillations, osteopathic medicine, health, upright posture

INTRODUCTION

In the osteopathic medicine field, therapeutics and clinical goals are focused on findings and treatments of somatic dysfunctions, which is described as an impaired or altered function of related components of the somatic (body framework) system, caused or expressed by any change of biological tissue structures in size, texture, structure, and position. However, until recently, the uncertainty linked to the pathophysiological processes leading to the onset of dysfunction and the lack of diagnostic reliability, based on tissue indicators, such as abnormalities in the cellular texture, asymmetries, restrictions of mobility, and greater tenderness or sensitivity,

make the concept of “osteopathic lesion” a subject of debate and further research (Liem, 2016). Hence, this study aimed to propose and develop a method for assessing the dysfunctional sphere, using knowledge of the communication points between the fascial system and the oscillatory adaptive capacity of the body in space to maintain an upright posture.

The physiology behind the upright posture in the human could appear as a static, stable, and imperturbable condition. The truth, however, seems to be much more complex and is still debated nowadays (Pollock et al., 2000). There is not yet a standard and universal model capable of correctly understanding and recognizing all the variables behind the physiology of postural control in humans.

For most adult individuals, the bipedal stance is a very simple and automatized task; however, the physiological processes inherent to this condition are quite complex and finalized to maintain and balance the body mass. This can be intrinsically unstable due to the gravitational load and the continuous perturbations caused by internal stimuli, such as heart beating and breathing, or external stimuli, including visual distortions or changes in the environment surrounding the individual (Forbes et al., 2018).

In order to maintain a stable balance in an unstable body, it is important to keep correct coordination and communication between the sensorineural system and musculoskeletal system, aiming to maintain the center of mass (COM) of the body inside the base of support (BoS) edges. This physiological function requires an unavoidable and necessary shift of the entire COM in a horizontal plane (Haddas and Lieberman, 2018). These continuous changes in the COM are expressed in the form of oscillations of the entire body, which, in physiological conditions, occur inside a stable region of the space surrounding the individual. This is called the cone of economy (COE; Dubousset, 1994), and it has a conical shape with the apex facing the BoS and the base facing the head. Any deviation and shifting from the center of the cone represent a challenge to keeping balance, requiring more muscle activation and energy expenditure (Savage and Patel, 2014).

Therefore, oscillations inside the COE represent individual dynamic skills in order to compensate perturbations caused by internal and external stimuli to maintain all those mechanical and physiological processes useful for maintaining balance and therefore the upright posture function.

The concept of stability and balance in human biomechanics is not well defined and universally recognized yet. However, it is possible to roughly adapt Newtonian mechanics principles to the biomechanical principles related to human balance physiology and the concepts behind the COE. In mechanical physics, the term “balance” is defined as the state of an object when the resultant load actions (forces or moments) acting upon it are zero (Newton’s First Law). In the case of unstable bodies, the more the line of gravity passing through the COM is distant from the center of BoS, the more precarious the balance will be. In inanimate objects, physical laws enable the fall of the object to be taken into consideration, if the line of gravity falls outside the BoS. The sensorineural and muscular postural control of the human body, on the other hand,

guarantees muscle activation forces against gravitational forces to ensure the maintenance of COM within the BoS (Pollock et al., 2000). These principles are reflected in the definition of COE given by Dubousset (1994) and in the concept of postural alignment, or ideal posture, which is defined as that state in which body mass is distributed in such a way that muscles tone, bones, and ligaments neutralize the force of gravity (Kappler, 1982).

In orthopedic surgery, postural models describe upright posture physiology as a complex relationship between the spine and pelvic morphology with axial and appendicular musculature. Their ideal alignment allows for the maintenance of a bipedal state with minimum energy expenditure (Savage and Patel, 2014). However, these models find it hard to consider important factors involved in upright postures, including the interdependence of body systems relative to a somatic framework, fascial mechanics, and bio-psycho-social dynamics of the individual (Lunghi et al., 2016).

In the osteopathic field, the “ideal posture” concept was defined by Kuchera (1995) as the result of dynamic interactions between individual force and gravity force. Posture, therefore, becomes an expression of the balance between these two forces and thus, an indicator of the adaptive capacity of an individual against the force of gravity. A recent review of the biomechanical model, defined as the integration between somatic components relative to the upright posture function, suggested that the “ideal posture” is a temporary and constantly dynamic result based on individual capacities expressed and integrated at a neuro-myofascial level and based on the interdependence of neural, somatic, and biopsychosocial systems (Lunghi et al., 2016).

HYPOTHESES

This study hypothesis aimed to analyze structural and organics elements involved in bipedal stance. They attempt to assess and possibly add to the equation the mechanics and the nature of the fascial system. Thinking about the concepts expressed by the COE and its relationships with the physiological mechanisms of posture, it would be appropriate to ask: Does a deal posture correspond to an ideal oscillatory sequence?

In the stance phase, body mass oscillations occur through fixed supports represented by the ankle and hip joints. Adaptations in body sways can occur on a sagittal plane, with anteroposterior oscillations, or on a frontal plane with mediolateral standing balance. During unperturbed postural conditions, the ideal oscillation patterns (ergonomic and with minimum energy expenditure) occur on a sagittal plane, with body sways equally distributed in an anteroposterior way (Forbes et al., 2018). The bone anatomy of the ankle joint, the relationships between the head of the talus and the tibiotarsal mortar and related soft tissue mechanics, seems to confirm to this predictive model. Only the geometry of this joint, together with a review of its biomechanics, confirms that, in the bipedal stance, range of movement is induced and facilitated primarily on a sagittal plane, providing resistance to eversion (Brockett and Chapman, 2016).

On the other hand, the knowledge of the authors about mediolateral standing balance in the frontal plane is much less clear and defined, as this also involves load/unload mechanisms by the hip abductors and adductors with a minimum counterbalancing activity from ankle invertors and evertors (Forbes et al., 2018).

This study aimed to assess individual capacity to discharge gravitational forces through body oscillations inside the COE, using the functional assessment of the body system based on the specificity of their structure. These functional evaluations should define “non-ideal” oscillations patterns, revealing postural perturbations showing a possible dysfunctional sphere. This leads the operator to follow a determined clinical rationale based on the knowledge and a treatment relative to the main body systems in the body, which include the musculoskeletal system, the visceral system, and the cranial system with their respective and specific techniques and treatment modalities. Besides internal and external stimuli cited before, further postural perturbation factors, acting upon body sways, could reside in the fascial system and its mechanics.

THE ROLE OF THE FASCIAL SYSTEM IN MAINTAINING AN UPRIGHT POSTURE

Several studies have shown the important role of the fascial system in controlling and maintaining an upright posture. Inherent myofascial tissue properties seem to play a decisive passive role, due to the tissue stiffness itself, in bipedal stance together with the active action of the neuromuscular and sensory systems. Therefore, these myofascial physiological properties occur without the control of the central nervous system. Indeed, as for the structure and histological anatomy of the myofascial tissue, it presents “tensegrative” properties capable to absorb and transmit gravity forces, generating a framework capable of supporting body mass and drastically reducing the neuromuscular effort required to maintain COM within the BoS. This inherent myofascial tissue property is called “Human Resting Myofascial Tone” and is integrated with the other passive body support systems, such as ligaments and bones (Masi et al., 2010).

The fascial system as connective tissue is rich in collagen fibers. It has been observed that the structural disposition of these fibers follows the main compression pathways mechanically defined when the entire body mass is under compressive forces, such as the force of gravity. Indeed, looking at the functional hierarchies of myofascial tissue, the epimysium presents bundles of collagen type I fibers disposed of longitudinally with respect to the sarcomere, forming a helical pattern arrangement at 55° relative to the longitudinal resting fusiform muscles axis. This angle is ideal for absorbing and transmitting both tensile and compressive forces (Scarr, 2016).

For example, lumbar multifidus has an inherent stiffness capable of stabilizing and maintaining an upright posture. Therefore, human resting myofascial tone can be considered a passive function of the myofascial tissue, due to its elastic

and stiffness properties. These specific properties occur only in a stationary phase, as viscoelastic properties of the fascial tissue tend to change during active movements. In short, during postural control, the fascia acts as a spring, without hysteresis, and is therefore useful for managing adaptive oscillations expressed inside the COE. This is a further confirmation of how the fascial system plays a decisive role in balancing and supporting an upright posture (Masi et al., 2010; Francavilla et al., 2020).

However, it can also be a source of perturbation and an additional force that acts on balancing body mass. Therefore, it is a variable to be added within the interceptive stimuli that act on postural control (Tozzi, 2015).

The fascial influence on posture has already been studied with significant benefits on postural assessment, followed by specific lengthening protocols of the main myofascial meridians (DellaGrotte et al., 2008).

Furthermore, there seems to be a possible communication between hypertonicity of the axial fascia and the onset of ankylosing spondylitis (Masi et al., 2010). Some fascia-related mechanisms seem to be associated with the somatic dysfunction model. For example, fascial ability to change its structure and texture-based upon environmental demands can increase stiffness and surrounding tensional states in soft tissues, with loss of elastic properties and the consequent inability of the tissues to properly absorb or transmit forces, such as gravity force (Tozzi, 2015). In view of these considerations, it seems reasonable to hypothesize that the COE should express higher amplitude oscillations toward the region of the human body with the most probable presence of somatic dysfunction. This phenomenon should occur as the result of the altered tensional state of the specific tissues, i.e., a passive, constant, and low-frequency postural perturbation disconnected from the central nervous system, but can easily be compensated through neuromuscular activations during postural control in an upright posture.

Curiously, similar anti-gravity mechanical properties have been observed not only in myofascial tissue related to the muscle skeletal system but also in connective tissue, which composes and gives shape and function to the visceral fascia and meningeal fascia (Nakagawa et al., 1994; Maikos et al., 2008; Stecco et al., 2017). Considering the evidence behind these tissue properties and the anatomical dispositions of the main fascial systems, evaluating dysfunctional oscillations patterns through a manual assessment of the COE is possible, guiding the manual therapist first toward a clinical approach, based on the proper rationale and techniques.

EVALUATION OF THE HYPOTHESIS

On summarizing, fascia is a ubiquitous viscoelastic tissue that forms a functional collagen matrix that surrounds and permeates all the body structures (somatic, visceral, neural, and vascular). This has made it much difficult, during years of research, to name and histologically isolate it. Many researchers have tried to define various types and nomenclatures of fascial tissues based on the relationship between their structures and their

functions; therefore, fascial nomenclature has been studied and debated. In order to assess the influence of the somatic dysfunction inside the COE, it is essential to know the relationship between the main fascial systems and their anatomical disposition around the human body framework.

The functional viscoelastic model, inherent to the myofascial tissue, is studied and described by Masi et al. (2010). It can be useful in a first partition and nomenclature of fascia. As an important anti-gravity and postural component, it is abundantly present in the body region posterior to the spine. As mentioned above, the lumbar multifidus is a perfect example.

The histological model described by Kumka and Bonar (2012) highlighted another functional fascial definition of clinical relevance. For example, the so-called “passive linking fasciae” seem to have a decisive role in passive force transmission. The latest evidence in research shows a specific fascial functioning with its histological mechanics, relative to that part of the somatic framework responsible for the upright posture function. On the other hand, the “separating fascia” system has the main function of compartmentalizing organs and body regions and maintaining their structural and functional integrity by promoting a correct sliding between viscera during movements and a good absorption and adaptation to mechanical stresses, as well as limiting the spread of infections (Liem, 2016).

The role of the latter anatomical structures on the maintenance of an upright posture is unclear and poorly studied. However, there are some histological and biomechanical considerations to take into account. The protective and compartmental function of the meningeal system on the central nervous system and spinal cord is universally recognized. It is also known that the human spine is stabilized not only by musculoskeletal components but also by soft tissues, through the fascia and spinal ligaments, both in the dynamic phase and in the stationary phase. Within these tissues, posterior longitudinal ligament (PLL) and spinal dura mater (SDM) are present. Evidence from histological examinations has outlined how the abundant presence of collagen and elastin fibers, arranged in a caudal sense, both in the PPL and in the ventral portion of the SDM, is extremely important in longitudinal mechanical stress responses. Furthermore, they provide resistance to tensions caused by the flexion of the spine, preventing SDM folding during hyperextension (Nakagawa et al., 1994). Also, the inherent connective tissue stiffness seems to play an important role in system stability.

Other histological research showed how SDM longitudinal fibers in the lumbar spine are 5–10 times denser than transversal fibers (Maikos et al., 2008). It seems like some of the functions inherent to soft tissue mechanics are linking, transferring, maintaining, and discharging longitudinal forces. Changes in SDM tensional states are necessary to alleviate postural stress and flexion/extension movements. Dural tissue acts in the same way as the collagen fascial tissues, through fibroblasts action, altering the surrounding textures.

The suboccipital region hosts an important anatomical component that connects the somatic framework with the dural system. It is called myodural bridge (MDB), and it is composed of fibrous connective tissue between the dura mater and the following:

1. Rectus capitis posterior minor (RCPmi).
2. Rectus capitis posterior major (RCPma).
3. Obliquus capitis inferior (OCI).
4. Nuchal ligament.

This system is characterized by bundles of fibers arranged parallel to each other in a longitudinal manner. It is assumed that the physiology of the MDB includes sensorimotor functions, postural control, and cerebrospinal fluid (CSF) flow modulation. This occurs by tensioning the SDM or transmitting the force generated by the movements of the head on the SDM itself (Zheng et al., 2018).

It is important to underline the relationship between the cranial sphere and meningeal fascia. Communication points between the pericranial musculature and the SDM are present, allowing muscles to modulate dural tensional states and vice versa. This hypothesis is supported by the presence of esocranial trigeminal nerve endings, which innervate the skull myofascial system (Bordoni et al., 2019a). Based on these observations, a sort of “connective pillar” is present and is represented by the spine and all the muscular, fascial, and connective elements which surround it. The inherent stiffness of these tissues should grant stability and functionality to the upright posture and support to the main anti-gravity systems.

The assessment of anti-gravity and postural function of fasciae should be evaluated through a diagnostic test called Pressure and Oscillation Test (POT), which is subdivided into two phases. The first phase consists of engaging the fascial anti-gravity function through a longitudinal pressure, applied by an operator on a subject from the vertex perpendicular to the ground. By absorbing and transmitting the lines of mechanical stress induced by the operator, the fascial systems accumulate potential energy waiting to be released (the same mechanical principle applicable to springs). By reducing the BoS of the subject, with feet together, a condition of instability is brought about in maintaining the balance of the body mass, stabilized temporarily by the caudal force applied by the operator. In the second phase, the operator rapidly stops the applied force by putting his hand out of the vertex of the subject. In this way, potential energy is suddenly released by the fascial and muscular tissues, in conjunction with closure of the eyes of the subject (**Figure 1**).

Changes in specific sensorial information, such as closure of the eyes, induce modulations of the whole-body postural oscillations system (Forbes et al., 2018). While the central nervous system is acting to settle a correct neuro-sensorial setting, which is necessary for maintaining an upright posture function, the body mass is under the influence of the force of gravity and passive tissues mechanics, which are disconnected from the central nervous system. At this exact moment, a facilitated oscillation toward the zone of main tissue tensions should emerge.

Considering the major postural instability induced by the reduction of the BoS, before the central nervous system is able to compensate for this condition through the correct activation of the neuromuscular system, it is possible to observe and assess the main oscillation emerging in that specific moment.

An increase of facilitation in swaying toward a specific region of the space inside the COE of an individual should

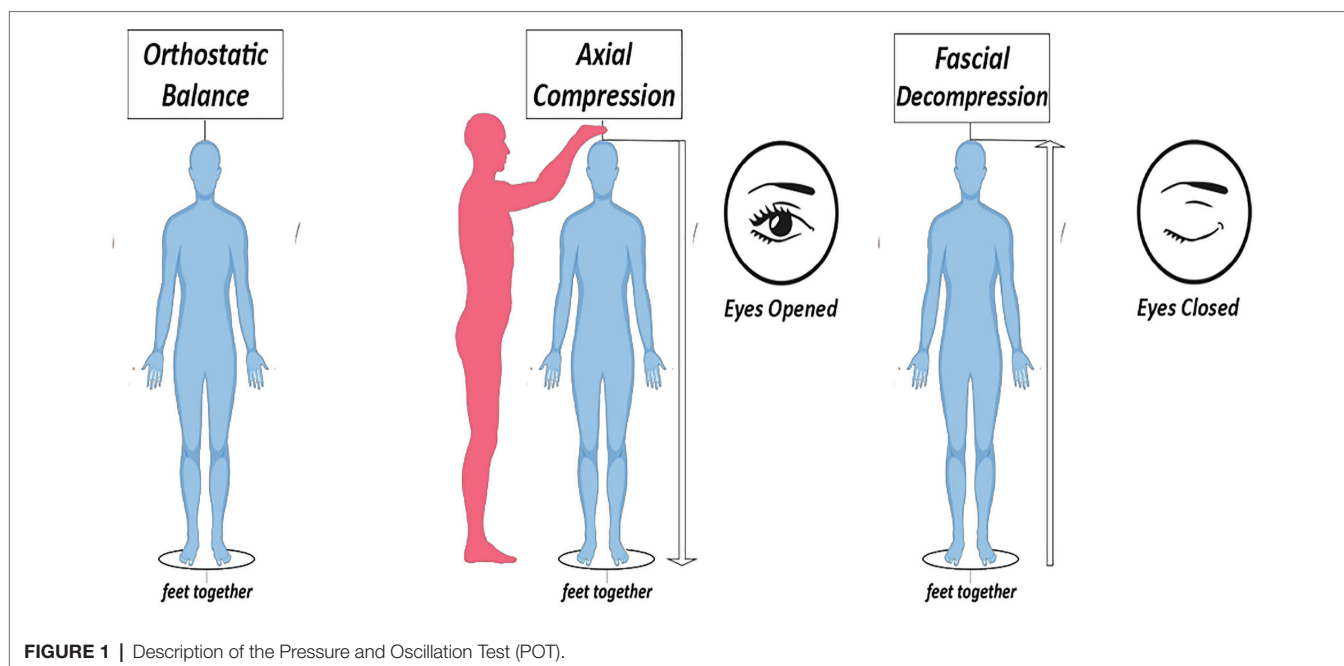


FIGURE 1 | Description of the Pressure and Oscillation Test (POT).

reveal the presence of more stiffness and loss of elasticity in fascial tissues. This would lead the therapist to assess the anatomical region, using first diagnostic and therapeutic approach.

DISCUSSION

Following this rationale, we were able to build a diagnostic model by determining dysfunctional oscillations patterns, using clinical and the histological (tissues mechanical) evidence behind the POT. A first dysfunctional indicator can be represented by a predominant posterior oscillation on a sagittal plane. This should bring the therapist to further assess the postural/muscle-skeletal/structural system, with the majority of the somatic components relative to this system located behind the “central pillar” of the human body, represented by the spine.

It was clinically observed that a major oscillation occurring in a frontal plane corresponds to alterations of palpable features relative to the meningeal system and its related fascial tissues. The evidence based on stabilometric data observations has shown how subjects affected by vestibular pathologies present increased variations of the center of pressure in a frontal plane compared with a control group (Nacci et al., 2011). The relationship between lateral oscillation and the craniosacral system is also described in the *Sacro Occipital Technique (SOT)* category II: “Category II involves over-motion or instability of the sacroiliac joint causing a dysfunctional relationship between the tailbone and pelvis. The sacroiliac weight-bearing joint sprain makes it difficult for the body to maintain itself against gravity and function at its optimum. It is common with this category to find low back imbalance, which can affect the knees, shoulders, neck and TMJ with Category II patients” (De Jarnette, 1984). One of the indicators of “category II” (sacroiliac weight-bearing dysfunction) described

by De Jarnette (1984) is the lateral sway of the entire spine either to the left or right, away from the center as seen on plumb line analysis (Hochman, 2005). This relationship allows researchers to relate the lateral oscillations on the frontal plane with the craniosacral system and, therefore, with the meningeal system. Further studies and insights in this regard are necessary.

On the other hand, anterior oscillations on a sagittal plane are characterized by changes in the structures of the visceral fascial systems classified as separating fascia 17, and are anatomically distributed behind the spine (Kumka and Bonar, 2012). The mechanics of visceral fascial tissues reflect the functionality analyzed so far to be typical of connective cells, with an extracellular matrix rich in elastin and parallelly arranged collagen bands capable of giving shape and structure to the internal organs and ensuring the transmission of forces. The exact type of innervation of many visceral bands is still uncertain. However, it is known that the parietal peritoneum has the same type of somatic innervation typical of the abdominal wall (Stecco et al., 2017).

Two different types of fascial tissue cover and give structure to each organ. A thin layer intimately adhering to each organ represents the first one, called investing fascia. It probably has an autonomic innervation and a prevalent elastic component. The second type, called insertional fascia, is thicker, fibrous, and multi-layered and functionally connects the various organs to the abdominal walls and to the musculoskeletal system. Between the various layers, there is a loose connective tissue that allows free sliding and motility to the respective organs. Therefore, alterations in tensional forces acting on visceral fascial tissues, through dysfunction in somatic components or an increase in tensions due to autonomic non-striated muscle cells, should act on and transmit directly to the postural somatic system (Bordoni et al., 2019b; **Figure 2**).

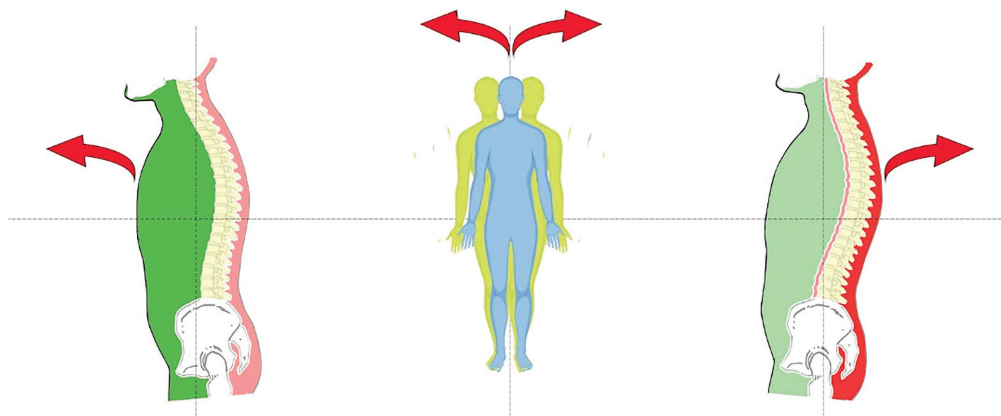


FIGURE 2 | Visceral Pattern (Green) – Meningeal/Cranial Pattern (Yellow) – Postural/muscle skeletal pattern (Red).

The evidence based on the physiology of the fascial tissues and on naturally expressed unstable balance through Newtonian mechanics support the clinical rationale behind the POT. A first evaluation and measurement method is underway to confirm the clinical validity inherent to the POT. This is done using a randomized clinical protocol that compares the biomechanical analyses of the COE of a group of subjects with respective manual tests of pressure and oscillations. A statistically significant outcome between the POT and the mechanical evaluation method would give substantial validity to the ability of the diagnostic test to express and broadly interpret the COE of a single individual. Future research should add a control group to compare a specific biomechanical analysis of the COE with oscillations observable exclusively through the closed eyes of the subjects.

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

AB, SC, GF, JP, SM, PA, and AR contributed to the conception and design of the study. NM, DD, and MC were responsible for drafting and finalizing the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

- Bordoni, B., Morabito, B., and Simonelli, M. (2019a). Cranial osteopathy: obscurantism and enlightenment. *Cureus* 11:e4730. doi: 10.7759/cureus.4730
- Bordoni, B., Simonelli, M., and Morabito, B. (2019b). The other side of the fascia: the smooth muscle part I. *Cureus* 11:e4651. doi: 10.7759/cureus.4651
- Brockett, C. L., and Chapman, G. J. (2016). Biomechanics of the ankle. *Orthop. Trauma* 30, 232–238. doi: 10.1016/j.mporth.2016.04.015
- De Jarnette, M. B. (1984). Sacro-occipital technique. Self Published; Nebraska City, NE.
- DellaGrotte, J., Ridi, R., Landi, M., and Stephens, J. (2008). Postural improvement using core integration to lengthen myofascia. *J. Bodyw. Mov. Ther.* 12, 231–245. doi: 10.1016/j.jbmt.2008.04.047
- Dubousset, J. (1994). “Three-dimensional analysis of the scoliotic deformity,” in *The Pediatric Spine: Principles and Practice*. ed. S. L. Weinstein (New York: Raven Press Ltd.).
- Forbes, P. A., Chen, A., and Blouin, J. S. (2018). Sensorimotor control of standing balance. *Handb. Clin. Neurol.* 159, 61–83. doi: 10.1016/B978-0-444-63916-5.00004-5
- Francavilla, V. C., Genovesi, F., Asmundo, A., Di Nunno, N. R., Ambrosi, A., Tartaglia, N., et al. (2020). Fascia and movement: the primary link in the prevention of accidents in soccer. Revision and models of intervention. *Med. Sport* 73, 291–301. doi: 10.23736/S0025-7826.20.03677-7
- Haddas, R., and Lieberman, I. H. (2018). A method to quantify the “cone of economy”. *Eur. Spine J.* 27, 1178–1187. doi: 10.1007/s00586-017-5321-2
- Hochman, J. I. (2005). The effect of sacro occipital technique category II blocking on spinal ranges of motion: a case series. *J. Manipulative Physiol. Ther.* 28, 719–723. doi: 10.1016/j.jmpt.2005.09.002
- Kappler, R. E. (1982). Postural balance and motion patterns. *J. Am. Osteopath. Assoc.* 81, 598–606.
- Kuchera, M. (1995). Gravitational stress, musculoligamentous strain, and postural alignment. *Spine State Art Rev.* 9, 463–489.
- Kumka, M., and Bonar, J. (2012). Fascia: a morphological description and classification system based on a literature review. *J. Can. Chiropr. Assoc.* 56, 179–191.
- Liem, T. A. T. (2016). Still’s osteopathic lesion theory and evidence-based models supporting the emerged concept of somatic dysfunction. *J. Am. Osteopath. Assoc.* 116, 654–661. doi: 10.7556/jaoa.2016.129
- Lunghi, C., Tozzi, P., and Fusco, G. (2016). The biomechanical model in manual therapy: is there an ongoing crisis or just the need to revise the underlying concept and application? *J. Bodyw. Mov. Ther.* 20, 784–799. doi: 10.1016/j.jbmt.2016.01.004
- Maikos, J. T., Elias, R. A., and Shreiber, D. I. (2008). Mechanical properties of dura mater from the rat brain and spinal cord. *J. Neurotrauma* 25, 38–51. doi: 10.1089/neu.2007.0348
- Masi, A. T., Nair, K., Evans, T., and Ghandour, Y. (2010). Clinical, biomechanical, and physiological translational interpretations of human resting myofascial tone or tension. *Int. J. Ther. Massage Bodywork* 3, 16–28. doi: 10.3822/ijtmb.v3i4.104
- Nacci, A., Ferrazzi, M., Berrettini, S., Panicucci, E., Matteucci, J., Bruschini, L., et al. (2011). Vestibular and stabilometric findings in whiplash injury and minor head trauma. *Acta Otorhinolaryngol. Ital.* 31, 378–389.
- Nakagawa, H., Mikawa, Y., and Watanabe, R. (1994). Elastin in the human posterior longitudinal ligament and spinal dura. A histologic and biochemical study. *Spine* 19, 2164–2169. doi: 10.1097/00007632-199410000-00006
- Pollock, A. S., Durward, B. R., Rowe, P. J., and Paul, J. P. (2000). What is balance? *Clin. Rehabil.* 14, 402–406. doi: 10.1191/0269215500cr342oa
- Savage, J. W., and Patel, A. A. (2014). Fixed sagittal plane imbalance. *Global Spine J.* 4, 287–296. doi: 10.1055/s-0034-1394126

- Scarr, G. (2016). Fascial hierarchies and the relevance of crossed-helical arrangements of collagen to changes in the shape of muscles. *J. Bodyw. Mov. Ther.* 20, 377–387. doi: 10.1016/j.jbmt.2015.09.004
- Stecco, C., Sfriso, M. M., Porzionato, A., Rumbold, A., Albertin, G., Macchi, V., et al. (2017). Microscopic anatomy of the visceral fasciae. *J. Anat.* 231, 121–128. doi: 10.1111/joa.12617
- Tozzi, P. (2015). A unifying neuro-fasciogenic model of somatic dysfunction—underlying mechanisms and treatment—part I. *J. Bodyw. Mov. Ther.* 19, 310–326. doi: 10.1016/j.jbmt.2015.01.001
- Zheng, N., Chi, Y. Y., Yang, X. H., Wang, N. X., Li, Y. L., Ge, Y. Y., et al. (2018). Orientation and property of fibers of the myodural bridge in humans. *Spine J.* 18, 1081–1087. doi: 10.1016/j.spinee.2018.02.006

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.

Copyright © 2021 Buscemi, Campisi, Frazzetto, Petrilligieri, Martino, Ambramo, Rapisarda, Maldonado, Di Corrado and Coco. 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.



A Canonical Set of Operations for Editing Dashboard Layouts in Virtual Reality

Tarek Setti and Adam B. Csapo *

Doctoral School of Multidisciplinary Engineering Sciences, Szechenyi Istvan University, Gyor, Hungary

OPEN ACCESS

Edited by:

Anna Esposito,
University of Campania "Luigi
Vanvitelli, Italy

Reviewed by:

Mihoko Niitsuma,
Chuo University, Japan
Javad Khodadoust,
Payame Noor University, Iran

*Correspondence:

Adam B. Csapo
csapo.adam@sze.hu

Specialty section:

This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

Received: 27 January 2021

Accepted: 28 June 2021

Published: 12 July 2021

Citation:

Setti T and Csapo AB (2021) A
Canonical Set of Operations for Editing
Dashboard Layouts in Virtual Reality.
Front. Comput. Sci. 3:659600.
doi: 10.3389/fcomp.2021.659600

Virtual reality (VR) is a powerful technological framework that can be considered as comprising any kind of device that allows for 3D environments to be simulated and interacted with via a digital interface. Depending on the specific technologies used, VR can allow users to experience a virtual world through their different senses, i.e., most often sight, but also through touch, hearing, and smell. In this paper, it is argued that a key impediment to the widespread adoption of VR technology today is the lack of interoperability between users' existing digital life (including 2D documents, videos, the Web, and even mobile applications) and the 3D spaces. Without such interoperability, 3D spaces offered by current VR platforms seem empty and lacking in functionality. In order to improve this situation, it is suggested that users could benefit from being able to create dashboard layouts (comprising 2D displays) for themselves in the 3D spaces, allowing them to arrange, view and interact with their existing 2D content alongside the 3D objects. Therefore, the objective of this research is to help users organize and arrange 2D content in 3D spaces depending on their needs. To this end, following a discussion on why this is a challenging problem—both from a scientific and from a practical perspective—a set of operations are proposed that are meant to be minimal and canonical and enable the creation of dashboard layouts in 3D. Based on a reference implementation on the MaxWhere VR platform, a set of experiments were carried out to measure how much time users needed to recreate existing layouts inside an empty version of the corresponding 3D spaces, and the precision with which they could do so. Results showed that users were able to carry out this task, on average, at a rate of less than 45 s per 2D display at an acceptably high precision.

Keywords: virtual reality, 3D spaces, 3D dashboards, *in-situ* spatial editing, digital cognitive artifacts

1 INTRODUCTION

The concept of virtual reality (VR) integrates all technologies that enable the simulation of three-dimensional environments, thereby creating an artificial spatial experience to its users (Kim, 2005; Neelakantam and Pant, 2017). Although VR is most often associated with immersive VR headsets—which are gaining popularity in a wide range of sectors, including the gaming industry, real estate, and manufacturing (Zhang et al., 2018)—even 3D desktop environments can be regarded as VR in their own right (Berki, 2020).

In the past few years, it has been often argued that VR has the potential to become a highly practical tool in daily life, as has been the case with smartphones and other ICT technologies, by

extending human capabilities and further contributing to the “merging” of human capabilities with those of increasingly intelligent ICT systems (Baranyi and Csapo, 2012; Baranyi et al., 2015; Csapo et al., 2018). The use of VR can thus be expected to far extend from its origins in “electronic gaming”, and to become widespread in hospitals, schools, training facilities, meeting rooms and even assisted living environments for the elderly (Lee et al., 2003; Riener and Harders, 2012; Slavova and Mu, 2018; Corregidor-Sanchez et al., 2020), transforming the common experience and effectiveness of humans in a variety of domains. Further, VR can help humans learn about and work in replicated environments which would be dangerous or hazardous under normal physical circumstances (Zhao and Lucas, 2015). Based on the above, several authors have come to the conclusion that VR can be seen as much more than a source of entertainment or a tool of visualization, and can in fact serve as a versatile infocommunication platform (Christiansson, 2001; Vamos et al., 2010; Galambos and Baranyi, 2011; Baranyi et al., 2015; Csapó et al., 2018; Pieskä et al., 2019).

On the other hand, it is necessary to point out that the widespread adoption of VR still faces some challenges, which mostly have to do with the practicality of the technology in terms of its ability to seamlessly integrate with users’ existing digital lives. Although VR has the potential to go further than just provide appealing visual experiences, by influencing how information is encountered, understood, retained and recalled by users (Csapó et al., 2018; Horvath and Sudar, 2018), this is only possible if a data-driven spatial experience is emphasized first and foremost.

Generally speaking, our investigation in this paper focuses on using desktop-based to integrate 2D digital content, since performing tasks using this kind of VR architecture carries benefits in terms of ease of access (widespread availability of laptops) and suitability for long periods of work (Lee and Wong, 2014; Berki, 2020; Bőczén-Rumbach, 2018; Berki, 2019). Specifically, the main problem that we discuss here is the use and re-configuration of dashboard layouts inside 3D spaces, i.e., how 2D display panels can be created and edited while a user is situated inside a 3D environment. As described in (Horváth, 2019), the evolution and increased use of 3D technologies brings with it the possibility of using applications that are significantly more powerful than 2D applications; however, 2D content remains a crucial part of 3D environments, enabling the integration of video, text, and web content (Horváth, 2019), rendering the environments useful for learning, researching, and even advertising (e.g., Berki, (2018) confirms that 2D advertisements displayed in 3D VR spaces have more effect on users than when they are displayed as banners on traditional webpages). Thus, we can ascertain that VR spaces provide a more seamless experience if existing 2D content can be integrated seamlessly into the 3D, and are amenable to spatial reconfiguration according to users’ needs. As we will see, this is not without challenges, hence we propose a set of operators and a workflow for the *in-situ* editing of dashboard layouts inside 3D spaces.

The paper is structured as follows. In **Section 2**, we provide an overview of existing desktop VR platforms and their use of

dashboard layouts for visualizing 2D content. This section also introduces the MaxWhere platform, which served as an environment for the reference implementation of our layout editing tool. In **Section 3**, a brief discussion is given on the scientific challenges behind *in-situ* editing capabilities for 3D spaces. Although we focus primarily on the placement and orientation of 2D displays, this discussion also applies more generally to editing the arrangement of 3D objects. In **Section 4**, a set of operations and a workflow is proposed for editing dashboards in 3D spaces. The operations are intended to be complete and minimal (i.e., canonical). Finally, in **Section 5**, we describe the results of an experimental evaluation based on the reconstruction of three different virtual spaces by 10 test subjects, and conclude that the proposed operations and workflow offer a viable approach to the editing of 2D content in 3D virtual spaces.

2 DASHBOARD LAYOUTS IN EXISTING VIRTUAL REALITY PLATFORMS

As mentioned earlier, desktop VR is gaining increasing traction in education and training environments, due to its capability of supplying real-time visualizations and interactions inside a virtual world that closely resembles its physical counterpart, at much lower cost than using real physical environments (Lee and Wong, 2014; Horvath and Sudar, 2018; Berki, 2020). A number of VR platforms can be used for such purposes. For example, Spatial.io enables people to meet through augmented or virtual reality, and to “drag and drop” files from their devices into the environment around them, with the aim of exchanging ideas and iterating on documents and 3D models *via* a 2D screen. Users can access Spatial.io meetings even from the web, just through a single click; then, it is possible to start working in a live 3D workspace by uploading 2D and 3D files.

Another example of a desktop VR platform well-suited to educational purposes is JanusXR / JanusWeb, a platform that allows users to explore the web in VR by providing VR-based collaborative 3D web spaces interconnected through so-called portals. JanusVR allows its users to create and browse spaces *via* an internet browser, client app or immersive displays, and to integrate 3D content, 2D displays (images, videos, and links), as well as avatars and chat features into the spaces. Users can navigate between spaces through so-called portals. Although Janus VR as a company is not longer active, the project lives on as a community-supported service.

A similar philosophy is reflected in the Mozilla Hubs service, which is referred to by its creator, Mozilla as a service for private social VR. This service is also accessible on a multitude of devices, including immersive headsets and the web browser. Spaces can be shared with others and collaboratively explored. 2D content can be dragged and dropped (or otherwise uploaded) into the spaces, and avatar as well as chat (voice and text) functionality is available.

Generally speaking, a common feature of such platforms when it comes to arranging 2D content is that the possible locations and poses for the 2D content are either pre-defined (i.e., when the user enters the space, the placeholders are already available into which

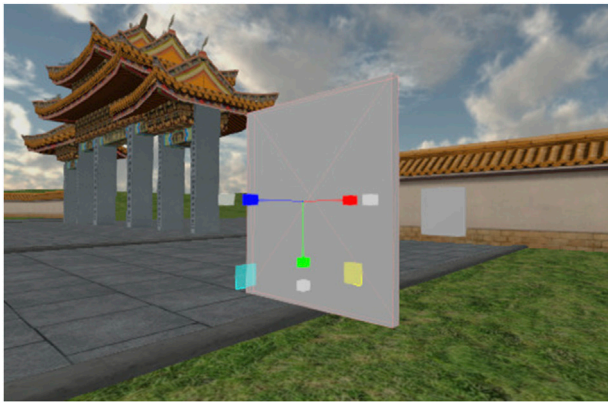


FIGURE 1 | An example of a gizmo transformation tool in JanusWeb.

the supported content types can be uploaded), and / or users are allowed to ‘drag and drop’ their local files into the 3D space, which then appears “in front” of the camera at any given time. In the latter case, the 2D display that is created inside the space is oriented in whatever direction the user is facing. In some cases, a geometric fixture (a “gizmo”) can appear overlaid on the object, which can be used as a transformation tool to move, rotate and re-size the object (**Figure 1**).

The proposed work in this paper is implemented on the MaxWhere desktop VR platform. MaxWhere was created with the purpose of enabling interaction with a large variety of digital content types (both 2D and 3D) in 3D spaces. MaxWhere can be accessed as a “3D browser” (a standalone client software compatible with Windows and MacOS) that can be used to navigate within and between 3D spaces available *via* a cloud backend. MaxWhere enables both interactive 3D models and 2D content (including videos, images, documents, webpages, and web applications, as well as social media platforms) to be integrated into the 3D spaces (see **Figure 2**). 2D content in MaxWhere is displayed in so-called “smartboards”, which integrate a Chromium-based rendering process and provide an interface similar to commonly used web browsers.

MaxWhere is already widely used in education, based on its capabilities for information sharing, visualization and simulation (Kuczmann and Budai, 2019; Rácz et al., 2020), therefore it is a strong candidate for replacing or at least supplementing traditional educational methods. It has been shown that users of MaxWhere

can perform certain very common digital workflows with 30 percent less user operations and up to 80 percent less “machine operations” (by providing access to workflows at a higher level of abstraction) (Horvath and Sudar, 2018), and in 50 percent less time (Lampert et al., 2018), due to the ability to display different content types at the same time. MaxWhere is also a useful tool for organizing virtual events such as meetings and conferences (more than one recent IEEE conference has been held on the MaxWhere platform, including IEEE Sensors 2020, and IEEE CogInfoCom 2020); and even for providing professional teams with an overview of complex processes in industrial settings (Bóczén-Rumbach, 2018).

The current version of MaxWhere does not include features for changing the arrangement of the 2D smartboards within spaces (although programmatically this is possible)—instead, the locations where users can display content are pre-specified by the designer of each space. Informally, many users the authors have spoken to have indicated that the ability to modify the arrangements would be a welcome enhancement to the platform. Accordingly, our goal is to propose a methodology that is universal and effectively addresses the shortcomings of the ‘gizmo approach’, which are described in **Section 3**.

3 CHALLENGES RELATED TO CREATING 2D LAYOUTS IN 3D

Despite the advantages of 3D spaces, both in general and in MaxWhere, a key obstacle in terms of integrating 2D content into 3D spaces is that users cannot change the existing 2D layouts (comfortably or at all), thus their thinking is forced into the constraints of the currently existing layouts, leading to reduced effectiveness and ease of use. However, it is also clear that no single layout is suitable for all tasks. In MaxWhere, it is often the case that users report their desire to add “just one more smartboard” to a space, or to “move this smartboard next to the other one” for clearer context.

In the case of platforms other than MaxWhere, where layout editing is enabled using the so-called “gizmo approach,” a new set of challenges become clear. When it comes to editing dashboard layouts in 3D, it is generally not a trivial question how the camera (the viewpoint of the user into the space) should interact with the operations used to transform the displays. If a display is being moved towards a wall or some other object, and the camera is in a stationary location, it will become difficult to determine when the



FIGURE 2 | 2D content inside MaxWhere 3D spaces.

display has reached a particular distance from the wall / object, and during rotation of the display, to determine whether the angle between the smartboard and the wall / object is as desired. However, if the camera viewpoint is modified automatically in parallel with the smartboard manipulations, users will be unable to re-position themselves with respect to the objects and smartboards of interest as freely as if the camera viewpoint is independent of the manipulations. This is a key dilemma, which we refer to as the “**camera-object independence dilemma**,” and which we have attempted to solve, in the dashboard layout editor proposed in this paper, by allowing the camera viewpoint to be independent of the operations, but helping to solve positioning challenges through features such as snapping the smartboards to objects, or smartboard duplication.

Another challenge in the design of an *in-situ* spatial editing tool is how to bridge between mathematical, geometric and physics-based concepts generally used by professionals in the design of 3D spaces (e.g., axes, angles, orientations expressed as quaternions, friction, gravitation, restitution etc.) and the terminologies that laypeople are better accustomed to (e.g., left/right/up/down/forward/backwards, visually symmetrical or asymmetrical, horizontal / vertical alignment, etc.). The key challenge is to design an interface that is intuitive to everyone, not just engineers, while allowing the same level of precision as would be required in the professional design of 3D spaces.

4 A CANONICAL SET OF OPERATIONS FOR EDITING 2D LAYOUTS

Currently, the 3D spaces available in MaxWhere each contain a space-specific default layout with smartboards having a pre-determined position, orientation, size, and default content. Therefore, in each space, users are forced to use the existing smartboards, even if their arrangement does not fulfill the users’ needs, which can be unique to any given user and application.

As we have indicated, the objective of this work is to help users organize and arrange dashboard layouts and 2D content in 3D spaces. This can be achieved by designing a set of operations intended to be canonical—that is, minimal (both in the sense of number of operations, and in the sense of number of operations it takes to achieve the same result) but at the same time also complete (allowing any layout to be created). The requirement of minimality necessitates the adoption of a workflow-oriented perspective, so that each individual 2D display can be placed into its final and intended pose based on the workflow and only based on the workflow. In this way, the length of the arrangement process (per display) can be quantified at a high level. In turn, the requirement of completeness can be evaluated experimentally, by asking a set of test subjects to re-create a variety of already existing dashboard layouts.

4.1 Workflow and Operations for Editing Layouts

Figure 3 provides a brief overview of the workflow process that was considered as a starting point as we developed the proposed canonical operations.

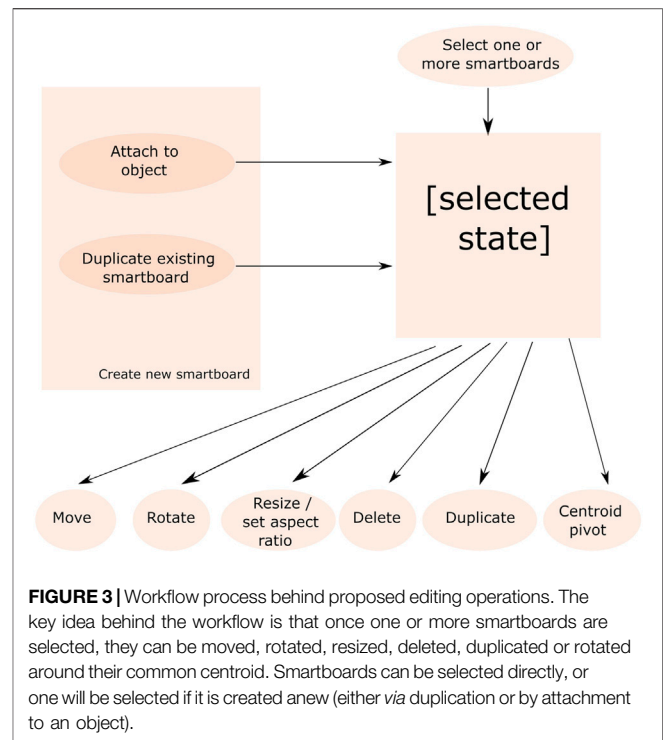


FIGURE 3 | Workflow process behind proposed editing operations. The key idea behind the workflow is that once one or more smartboards are selected, they can be moved, rotated, resized, deleted, duplicated or rotated around their common centroid. Smartboards can be selected directly, or one will be selected if it is created anew (either via duplication or by attachment to an object).

As detailed in the figure, the first step in editing a layout is to either add a new smartboard and then select it for further manipulations, or to select an existing smartboard for further manipulations. It is necessary to select one or more smartboards before it becomes possible to modify them. By “modification,” we mean some combination of scaling, translation and rotation operations. The different operations can be described as follows:

- 1) The first operation, in the case where a new smartboard is added to the scene, consists of choosing a place and an orientation to add the smartboard (**Figure 4**). Given that orientations are difficult to provide explicitly, the user is required to point the 3D cursor to a point on a surface inside the space (we refer to this point as the “point of incidence”), then exit (with the click of the right mouse button) to the 2D menu. This will “freeze” the spatial navigation and the user can click on an Add icon on the menu to create the new smartboard. The editor component then creates the new smartboard, assigns a unique integer ID to it, and places it in the position and orientation determined by the point of incidence on the surface (*via* the normal vector of the surface at that same point, as shown on **Figure 4**).
- 2) The second operation is that of selection, which allows users to select either individual smartboards, or groups of smartboards for further manipulation (**Figure 5**). Smartboards can be selected by toggling the selection mode (depicted by a “magic wand” icon in the 2D menu) and returning to 3D navigation mode, then hovering the 3D cursor over the given smartboard or smartboards. One can deselect individual or multiple smartboards by hovering the 3D cursor over them once again, or by toggling the selection mode again.



FIGURE 4 | Placement and addition of a new smartboard into the scene can be done by clicking the right mouse button to open the 2D “pebbles” menu and by clicking on the red plus sign.

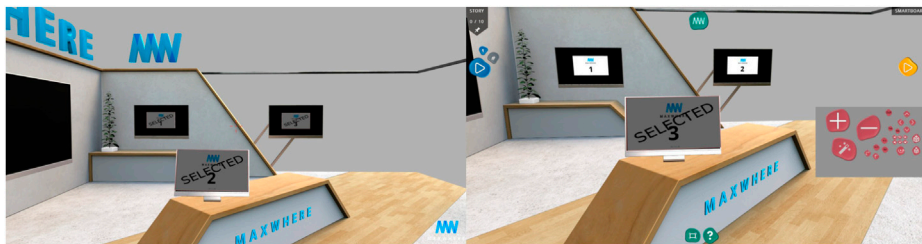


FIGURE 5 | Smartboard selection is performed by first toggling the selection mode (by clicking the magic wand among the red pebbles) and by hovering the mouse over the given smartboards in 3D navigation mode.

3) Once an individual smartboard, or a group of smartboards are selected, it / they can be further manipulated in terms of their position, orientation, size, and aspect ratio.

- One trivial manipulation is to delete the smartboard.
- Another is to duplicate it (in case only a single smartboard is selected). In the case of duplication, a newly created smartboard will be placed next to the originally selected smartboard).
- When it comes to manipulating position and orientation, the axes along which translation / rotation is performed depends on how many smartboards are selected. If a single smartboard is selected, it can be translated / rotated along or around its local axis (left / right axis, up-down axis and forward-backwards axis). If more than a single smartboard is selected and the smartboards are therefore being manipulated at the same time, the axes of manipulation become aligned with the global axes of the space. The reason for this is that when more than a single smartboard is selected, they may be facing different orientations, which means there is no single local axis for the left/right, up/down and forward/backwards directions. Further, if each smartboard were to be transformed along a different axis, the relative arrangement, and orientation of the smartboards would change in unexpected ways.
- When it comes to setting the aspect ratio of one or more smartboards, three different aspect ratios can be chosen: 16-by-9, 4-by-3, and A4 size. In each case, the width of each smartboard is kept, and their height is modified according to the chosen aspect ratio. It is also possible to scale the

smartboards to make them larger or smaller using two separate icons on the editor menu.

- Finally, an interesting but useful feature in the editor menu is the “centroid pivot” manipulation operation, which allows a group of smartboards to be rotated around their center of geometry. This enables users to mirror complete configurations (groups of smartboards) in different corners of the space, without having to move and rotate them individually, or to recreate the whole arrangement in a different location (**Figure 6**).
- 4) One experimental but quite useful feature added to the editor is the “snap-to” feature. As mentioned earlier, when a new smartboard is created, it is already added to the space in such a way that its orientation corresponds to the normal vector of the surface at the point where it is added to the space. However, users may later decide to move those smartboards to different locations, or incidentally the surface may not be completely flat. In such cases, users can temporarily turn on a physics simulation (by pressing Ctrl + S for “control + snap”—or Cmd + S on Mac OS), such that a gravity vector is added to the given smartboard, with the direction of the gravity vector pointing in the backwards direction. In this way, any smartboard can be forced to gravitate towards and snap onto any surface other than the one it was originally placed on.
- 5) At any point during the editing process, users may want to undo the previously effected operations. This is possible by pressing Ctrl + U (or Cmd + U on Mac OS). In our implementation, given the high precision of the

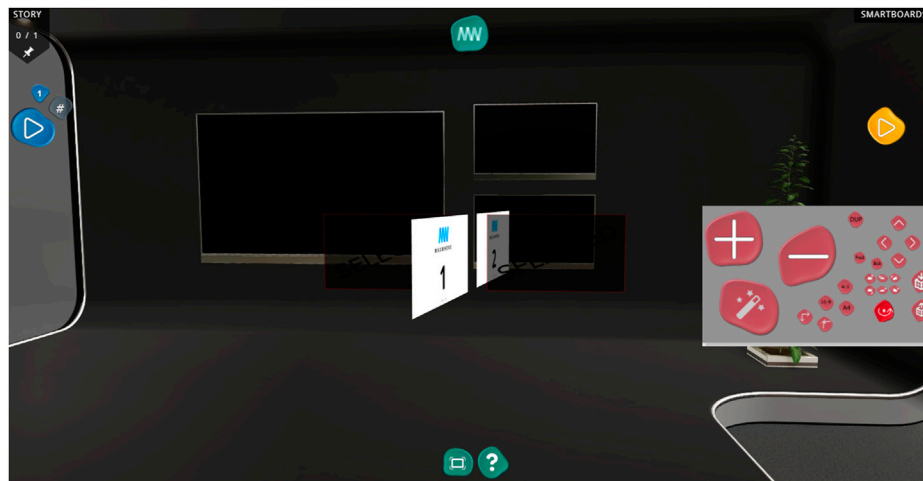


FIGURE 6 | Using the centroid pivot, it is possible to rotate a group of smartboards around their center of geometry. In this case, the original location of the two smartboards is indicated by the greyed out rectangles. As can be seen, both smartboards have been rotated towards the left around their center of geometry.

manipulations (see also **Section 4.2** below), we decided to give users the option of undoing the last 150 operations.

- 6) Finally, the export / import icons allow users to save (in .json format) and load the layouts, and to share those layouts with others.

4.2 Precision of the Manipulation Operations

As mentioned earlier in **Section 3**, the question of how the editing operations can be made precise without resorting to professional concepts remains a challenge. In our implementation of the editor tool, we facilitated precision by setting the gradations of manipulation to low values (0.5 cm in the case of translation and scaling, and 0.5 in the case of rotation). In order to ensure that the manipulation of the scene does not become overly tedious, we also implemented the manipulation buttons such that their sensitivity be rate-based. Thus, when clicked at a low rate, the small gradation values apply. When clicked at higher, more rapid rates, each gradation value is increased dynamically (up to 50 cm per click in the case of translation and scaling, and up to five degrees in the case of rotation).

5 EXPERIMENTAL EVALUATION AND DISCUSSION

As described earlier, the objective of this work was to design a set of operations that are complete and minimal (i.e., canonical) and enable users to reconfigure dashboard layouts inside 3D spaces. As detailed earlier, our goal was also to propose a set of operations that effectively address the camera-object independence dilemma and are interpretable not only to professional, but also to less seasoned users.

To verify the viability and efficacy of our tool, we realized an experiment with 10 test subjects. Each of the test subjects were

students at Széchenyi István University (six male, and four female test subjects between ages 20 and 35). The test subjects had varied experience with VR technologies and MaxWhere in particular; some had previously encountered MaxWhere during their studies at the university, while others were completely new to the platform. In the experiment, test subjects were given the task of re-creating the smartboard layouts in three different MaxWhere spaces from scratch (based on images of an original layout in each case, as shown in **Figure 7**). Following each task, the resulting layout was exported from the space, and the time it took to re-create the layout and the precision with which the task was accomplished was evaluated.

5.1 Experiment on Speed and Precision

In our experiment, we set out to measure the time it took the same 10 users to recreate the smartboards inside an empty version of three different 3D spaces, based on screenshots of the original layout of the same spaces. The three spaces used in the experiment were “Glassy Small” (13 smartboards), “Let’s Meet” (10 smartboards) and “Seminar on the Beach” (12 smartboards). In each case (space per user), we measured the time it took to complete the task, and compared the original layouts to the re-created ones by exporting the re-created layouts.

As shown on **Figure 8**, all users were able to finish the layout editing task in less than 20 min, with the average time being more like 10 min. This means that, on average, users required around 45 s per smartboard when re-creating the layouts. This seems acceptable for users new to the tool; however, in a second part of our analysis, we also wanted to evaluate the precision with which they were able to accomplish the task.

The precision per test subject and per space in terms of position, orientation, aspect ratio and smartboard area is shown in **Figure 9**. Details on the analyses are as follows:

- Difference in position was measured as the distance between the center of each smartboard and the re-created version of



FIGURE 7 | Three spaces at the center of the experiment (from top to bottom: “Glassy Small”, “Let’s Meet” and “Seminar on the Beach”).

the same smartboard, in centimeters (the basic distance unit in MaxWhere). To aggregate these values, the root mean squared value of the distances was calculated.

- Difference in orientation was measured as the difference in radians between the orientation of each smartboard and its re-created version. To aggregate these values, the root mean squared value of the differences was calculated.

Note that orientation in MaxWhere is specified using quaternions, that is, 4-dimensional vectors in which the x,y,z coordinates represent the coordinates of a global axis of rotation (scaled by the sine of half the angle of rotation) and in which the “w” coordinate corresponds to the cosine of half the angle of

rotation. Since any given quaternion corresponds to a specific rotation around a specific axis, it can be interpreted as a global orientation that is arrived at if the object is transformed by that rotation from its default state (the default state being the orientation obtained when the axes of the object are aligned with the global x,y,z axes of the space). The ‘distance’ (in radians) between two quaternions, q_1 and q_2 , then, can be expressed as follows (assuming that the norm of both quaternions is 1):

$$\theta = \arccos(2 \langle q_1, q_2 \rangle^2 - 1) \quad (1)$$

where $\langle q_1, q_2 \rangle$ is the scalar product of the two quaternions (summed product of coordinates in each corresponding dimension).

This can be shown to be true based on the fact that the scalar product of two unit-length quaternions always yields the cosine of half the angle between them, and the double-angle formula for cosines as follows:

$$\begin{aligned} \cos(0.5\theta) &= \langle q_1, q_2 \rangle \\ \cos(0.5\theta)^2 &= \langle q_1, q_2 \rangle^2 \end{aligned}$$

and, since $\cos(2x) = \cos(x)^2 - \sin(x)^2 = 2\cos(x)^2 - 1$ (also making use of the fact that $\sin(x)^2 + \cos(x)^2 = 1$), substituting 0.5θ for x , we have that:

$$\begin{aligned} \cos(\theta) &= 2\cos(0.5\theta)^2 - 1 \\ &= 2 \langle q_1, q_2 \rangle^2 - 1 \end{aligned}$$

- Difference in aspect ratio was measured as the difference between the (dimensionless) value of *width/height* for each smartboard and the re-created version of the same smartboard. To aggregate these values, the root mean squared value of the differences was calculated.
- Finally, differences in the surface areas of the original and re-created smartboards were also calculated on a pairwise basis and aggregated using the root mean squared value.

5.2 Interpretation of Experimental Results

Based on the results shown in Figures 8, 9, we were able to conclude that the proposed operations were complete (i.e., the layouts of very different spaces could be re-created) and effective (the users were able to complete the task in an acceptable amount of time).

Looking more closely at the precision results, one interesting observation was that error rates were significantly higher in the case of the “Glassy Small” space for all users, whereas the layout of the “Let’s Meet Extra” space was reconstructed most easily. This should be no surprise, given that the latter space—and to some extent the “Seminar Beach” space as well—featured visual placeholders (TV screen meshes or other frames)—which helped users in the positioning and sizing of the smartboards. Unsurprisingly, the discrepancy among the three spaces was smallest in the case of orientation, since this was the only measured aspect that was independent of the presence of such placeholders.

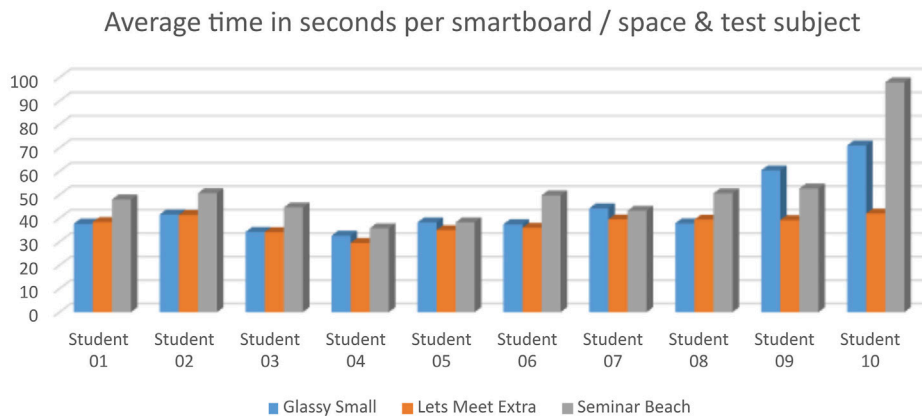


FIGURE 8 | Average time required to re-create the layouts of existing spaces was found to be less than 45 s / smartboard.

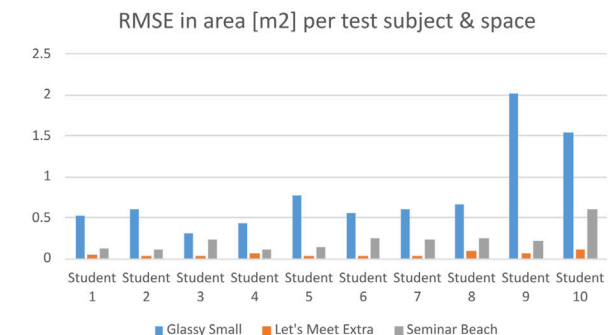
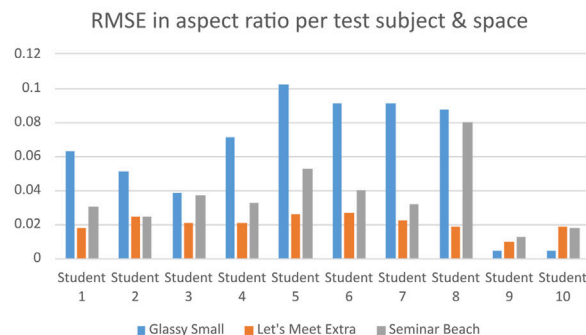
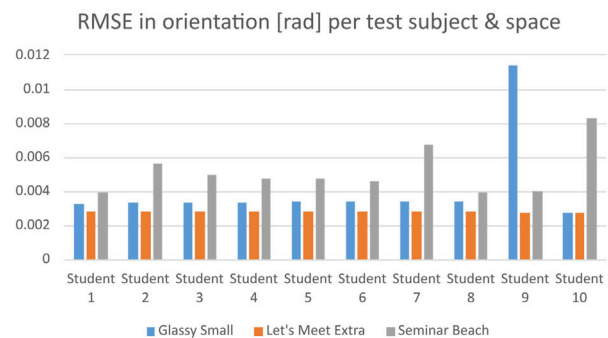
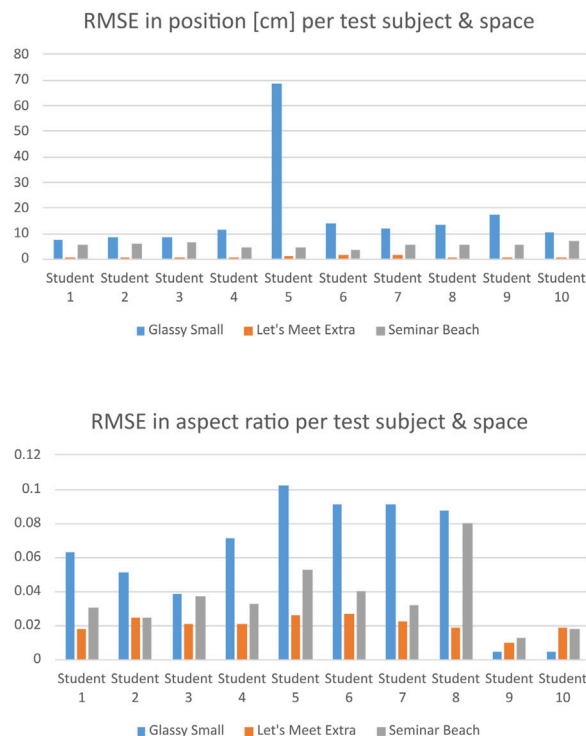


FIGURE 9 | Precision of the manipulation operations, measured as root mean squared error in centimeters (for position), radians (for orientation), and meters squared (for surface area).

Even so, in the case of the “Glassy Small” space, the error in smartboard positioning was generally around 10 cm—an insignificant error considering that each original smartboard in the space had a width of at least 1.40 m. Further, the high precision of reconstruction achieved in the case of the other two spaces suggests that when there were sufficient visual cues present, the challenges of the task were more than manageable.

In scenarios where it is up to the user to define his or her own dashboard layouts (without a given example to be re-created), the

proposed operations and workflow seem to be both efficient and effective.

6 CONCLUSION

In this paper, we argued that the widespread adoption of VR technologies in everyday use faces challenges in terms of the current ability of VR to integrate users’ existing digital life (e.g.,

2D documents, images, and videos) with 3D objects. To address this challenge, we focused on the ability of users to create dashboard layouts of 2D content while they are inside a 3D space. After identifying key challenges associated with current object manipulation methods (e.g. the “gizmo approach”)—namely the camera-object independence dilemma and the problem of interpretability of manipulation operations for end users, we proposed a workflow and an associated set of operations for the *in-situ* creation and manipulation of dashboard layouts in VR spaces. We validated the proposed methodology in terms of efficiency and completeness by having test subjects re-construct existing dashboard layouts in empty versions of otherwise existing VR spaces on the MaxWhere platform.

REFERENCES

- Baranyi, P., and Csapó, Á. (2012). Definition and Synergies of Cognitive Infocommunications. *Acta Polytech. Hungarica*. 9, 67–83.
- Baranyi, P., Csapo, A., and Sallai, G. (2015). *Cognitive Infocommunications (CogInfoCom)*. Springer.
- Berki, B. (2018). 2d Advertising in 3d Virtual Spaces. *Acta Polytech. Hungarica*. 15, 175–190. doi:10.12700/aph.15.3.2018.3.10
- Berki, B. (2019). Does Effective Use of Maxwhere Vr Relate to the Individual Spatial Memory and Mental Rotation Skills?. *Acta Polytech. Hungarica*. 16, 41–53. doi:10.12700/aph.16.6.2019.6.4
- Berki, B. (2020). “Level of Presence in max where Virtual Reality,” in 2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Mariehamn, Finland, 23–25 Sept. 2020 (IEEE), 000485–000490.
- Bóczén-Rumbach, P. (2018). “Industry-oriented Enhancement of Information Management Systems at Audi Hungaria Using Maxwhere’s 3d Digital Environments,” in 2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Budapest, Hungary, 22–24 Aug. 2018 (IEEE), 000417–000422.
- Christiansson, P. (2001). “Capture of user requirements and structuring of collaborative VR environments,” in Conference on Applied Virtual Reality in Engineering and Construction Applications of Virtual Reality, Gothenburg, October 4–5, 2001, 1–17.
- Corregidor-Sánchez, A.-I., Segura-Fragoso, A., Criado-Álvarez, J.-J., Rodríguez-Hernández, M., Mohedano-Moriano, A., and Polonio-López, B. (2020). Effectiveness of Virtual Reality Systems to Improve the Activities of Daily Life in Older People. *Ijerp* 17, 6283. doi:10.3390/ijerp17176283
- Csapó, Á. B., Horvath, I., Galambos, P., and Baranyi, P. (2018). “Vr as a Medium of Communication: from Memory Palaces to Comprehensive Memory Management,” in 2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Budapest, Hungary, 22–24 Aug. 2018 (IEEE), 000389–000394.
- Galambos, P., and Baranyi, P. (2011). “Virca as Virtual Intelligent Space for Rt-Middleware,” in 2011 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), Budapest, Hungary, 3–7 July 2011 (IEEE), 140–145.
- Horváth, I. (2019). Maxwhere 3d Capabilities Contributing to the Enhanced Efficiency of the Trello 2d Management Software. *Acta Polytech. Hungarica*. 16, 55–71. doi:10.12700/aph.16.6.2019.6.5
- Horvath, I., and Sudar, A. (2018). Factors Contributing to the Enhanced Performance of the Maxwhere 3d Vr Platform in the Distribution of Digital Information. *Acta Polytech. Hungarica*. 15, 149–173. doi:10.12700/aph.15.3.2018.3.9
- Kim, G. (2005). *Designing Virtual Reality Systems*. Springer.
- Kuczmann, M., and Budai, T. (2019). Linear State Space Modeling and Control Teaching in Maxwhere Virtual Laboratory. *Acta Polytech. Hungarica*. 16, 27–39. doi:10.12700/aph.16.6.2019.6.3
- Lampert, B., Pongracz, A., Sipos, J., Vehrer, A., and Horvath, I. (2018). Maxwhere Vr-Learning Improves Effectiveness over Classical Tools of E-Learning. *Acta Polytech. Hungarica*. 15, 125–147. doi:10.12700/aph.15.3.2018.3.8
- Lee, E. A.-L., and Wong, K. W. (2014). Learning with Desktop Virtual Reality: Low Spatial Ability Learners Are More Positively Affected. *Comput. Educ.* 79, 49–58. doi:10.1016/j.compedu.2014.07.010
- Lee, J. H., Ku, J., Cho, W., Hahn, W. Y., Kim, I. Y., Lee, S.-M., et al. (2003). A Virtual Reality System for the Assessment and Rehabilitation of the Activities of Daily Living. *CyberPsychology Behav.* 6, 383–388. doi:10.1089/109493103322278763
- Neelakantam, S., and Pant, T. (2017). *Learning Web-Based Virtual Reality: Build and Deploy Web-Based Virtual Reality Technology*. (Apress).
- Pieskä, S., Luimula, M., and Suominen, T. (2019). Fast Experimentations with Virtual Technologies Pave the Way for Experience Economy. *Acta Polytech. Hungarica*. 16, 9–26. doi:10.12700/aph.16.6.2019.6.2
- Rácz, A., Gilányi, A., Bólya, A. M., Décei, J., and Chmielewska, K. (2020). “On a Model of the First National Theater of Hungary in Maxwhere,” in 2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Mariehamn, Finland, 23–25 Sept. 2020 (IEEE), 000573–000574.
- Riener, R., and Harders, M. (2012). “Vr for Medical Training,” in *Virtual Reality in Medicine* (Springer), 181–210. doi:10.1007/978-1-4471-4011-5_8
- Slavova, Y., and Mu, M. (2018). “A Comparative Study of the Learning Outcomes and Experience of Vr in Education,” in 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Tuebingen/Reutlingen, Germany, 18–22 March 2018 (IEEE), 685–686.
- Vamos, A., Fülöp, I., Resko, B., and Baranyi, P. (2010). Collaboration in Virtual Reality of Intelligent Agents. *Acta Electrotechnica et Informatica* 10, 21–27.
- Zhang, M., Zhang, Z., Chang, Y., Aziz, E.-S., Esche, S., and Chassapis, C. (2018). Recent Developments in Game-Based Virtual Reality Educational Laboratories Using the Microsoft Kinect. *Int. J. Emerg. Technol. Learn.* 13, 138–159. doi:10.3991/ijet.v13i01.7773
- Zhao, D., and Lucas, J. (2015). Virtual Reality Simulation for Construction Safety Promotion. *Int. J. Inj. Control Saf. Promot.* 22, 57–67. doi:10.1080/17457300.2013.861853

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

Both authors contributed 50% to the paper. TS worked on the conceptualization and implementation of the tool and carried out the experiments. AC also worked on the conceptualization and implementation of the tool, and helped design the experiments. Most of the manuscript was written by TS and proof-read by AC.

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.

Copyright © 2021 Setti and Csapo. 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.



Robot Concept Acquisition Based on Interaction Between Probabilistic and Deep Generative Models

Ryo Kuniyasu¹, Tomoaki Nakamura^{1*}, Tadahiro Taniguchi² and Takayuki Nagai^{3,4}

¹Department of Mechanical Engineering and Intelligent Systems, The University of Electro-Communications, Tokyo, Japan,

²College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan, ³Department of Systems Innovation, Graduate School of Engineering Science, Osaka University, Osaka, Japan, ⁴Artificial Intelligence EXploration Research Center, The University of Electro-Communications, Tokyo, Japan

OPEN ACCESS

Edited by:

Anna Esposito,
University of Campania "Luigi
Vanvitelli, Italy

Reviewed by:

Attila Kovari,
University of Dunaújváros, Hungary
Juan Sebastian Olier,
Tilburg University, Netherlands

*Correspondence:

Tomoaki Nakamura
tnakamura@uec.ac.jp

Specialty section:

This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

Received: 26 January 2021

Accepted: 06 August 2021

Published: 03 September 2021

Citation:

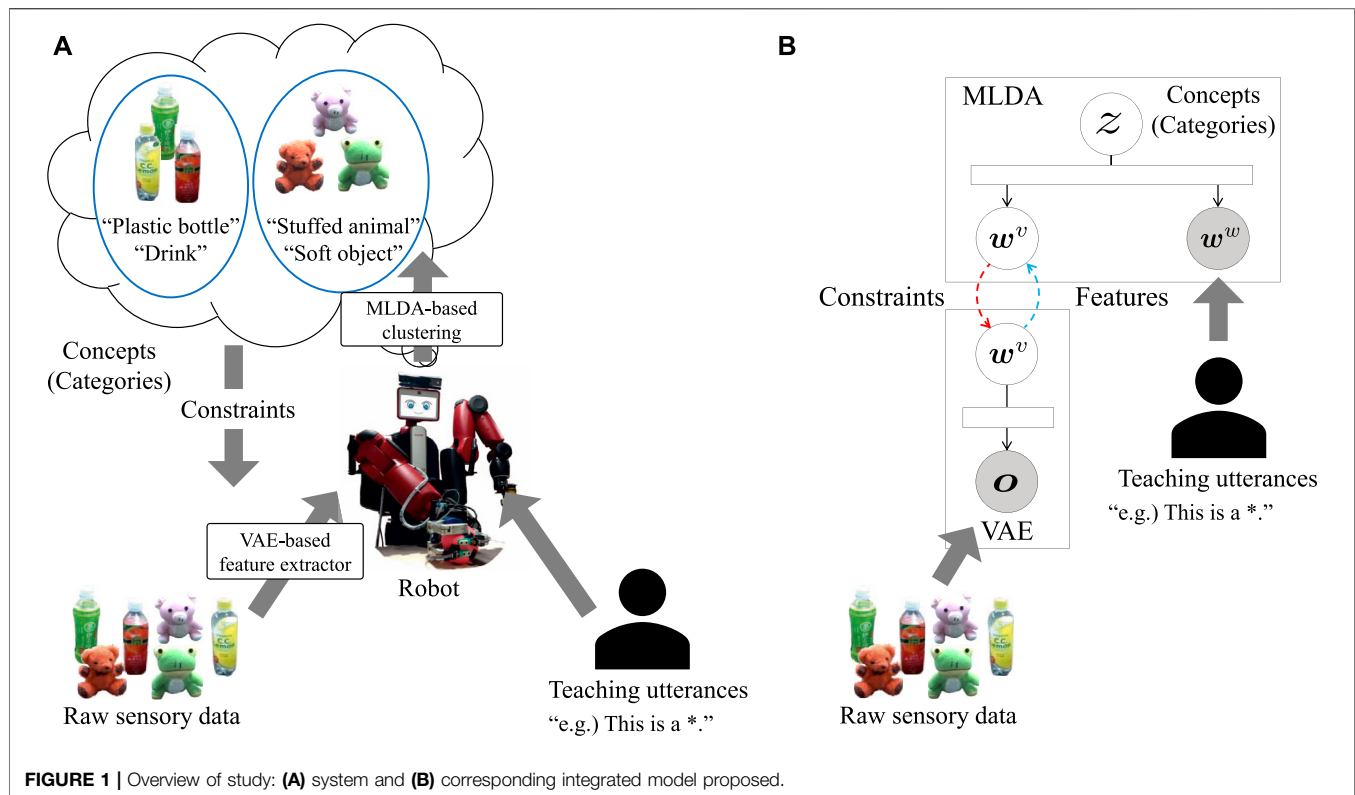
Kuniyasu R, Nakamura T, Taniguchi T
and Nagai T (2021) Robot Concept
Acquisition Based on Interaction
Between Probabilistic and Deep
Generative Models.
Front. Comput. Sci. 3:618069.
doi: 10.3389/fcomp.2021.618069

We propose a method for multimodal concept formation. In this method, unsupervised multimodal clustering and cross-modal inference, as well as unsupervised representation learning, can be performed by integrating the multimodal latent Dirichlet allocation (MLDA)-based concept formation and variational autoencoder (VAE)-based feature extraction. Multimodal clustering, representation learning, and cross-modal inference are critical for robots to form multimodal concepts from sensory data. Various models have been proposed for concept formation. However, in previous studies, features were extracted using manually designed or pre-trained feature extractors and representation learning was not performed simultaneously. Moreover, the generative probabilities of the features extracted from the sensory data could be predicted, but the sensory data could not be predicted in the cross-modal inference. Therefore, a method that can perform clustering, feature learning, and cross-modal inference among multimodal sensory data is required for concept formation. To realize such a method, we extend the VAE to the multinomial VAE (MNVAE), the latent variables of which follow a multinomial distribution, and construct a model that integrates the MNVAE and MLDA. In the experiments, the multimodal information of the images and words acquired by a robot was classified using the integrated model. The results demonstrated that the integrated model can classify the multimodal information as accurately as the previous model despite the feature extractor learning in an unsupervised manner, suitable image features for clustering can be learned, and cross-modal inference from the words to images is possible.

Keywords: concept formation, symbol emergence in robotics, probabilistic generative model, deep generative model, unsupervised learning, representation learning, cross-modal inference

1 INTRODUCTION

Not only multimodal clustering and cross-modal inference but also representation learning is critical for robots to form multimodal concepts from sensory data. We define multimodal categories that enable multimodal clustering and cross-modal inference as concepts. Cross-modal inference enables a robot to infer unobserved information from limited observed information, and this ability allows robots to respond to uncertain environments. Various models have been proposed for concept formation (Nakamura et al., 2014; Taniguchi et al., 2017). These models perform clustering and cross-modal inference based on the multimodal latent Dirichlet allocation (MLDA) (Nakamura



et al., 2009). However, pre-designed or pre-trained feature extractors are used, and representation learning is not performed simultaneously. Feature learning is an ability required for robots to obtain environment-dependent knowledge and adapt to the environment. Moreover, in these previous studies, the generative probabilities of the features extracted from sensory data could be predicted, but the sensory data could not be predicted in the cross-modal inference. Therefore, a method that can perform clustering, feature learning, and cross-modal inference of the sensory data is required for concept formation.

In this study, we focus on concept formation from image and word information, and propose a method to perform clustering and to learn suitable feature extractors simultaneously by integrating the MLDA-based concept formation and variational autoencoder (VAE)-based (Kingma and Welling, 2013) feature extractor. **Figure 1** presents an overview of this study. A robot forms concepts from the images captured by the camera and words given by the human user who teaches the object features. Furthermore, the formed concepts influence the learning of the feature extractor, making it possible to extract suitable features for concept formation in the environment.

Intelligent robots that learn from the information obtained from the environment and the humans (Ridge et al., 2010; Taniguchi et al., 2010; Mangin et al., 2015; Wächter and Asfour, 2015) have been developed in the field known as symbol emergence in robotics (Taniguchi et al., 2016, 2018). This field includes various research topics, such as concept formation (Nakamura et al., 2007; Ridge et al., 2010; Mangin

et al., 2015), language acquisition (Mochihashi et al., 2009; Neubig et al., 2012), learning of interaction (Taniguchi et al., 2010), and segmentation and learning of actions (Wächter and Asfour, 2015; Nagano et al., 2019). The symbol emergence system was proposed by Taniguchi (Taniguchi et al., 2016, 2018), and it was inspired by the genetic epistemology proposed by Piaget (Piaget and Duckworth, 1970). In the symbol emergence system, robots self-organize into symbols from the multimodal sensory information obtained from the environment in a bottom-up manner. In this case, bottom-up means learning knowledge only from the sensory information obtained from the environment, without hand-crafting it or training it with artificial disambiguated information such as teacher signals and labels, which cannot be used in the human learning process. Furthermore, the symbols (e.g., language) are shared with others and the self-organization is influenced by the shared symbols through communication with others in a top-down manner. The symbols emerge from these bottom-up and top-down loops.

We believe that it is important for robots to learn such symbols autonomously and in an unsupervised manner in this loop to coexist with humans. Accordingly, we have proposed models that allow robots to acquire concepts and language by clustering the multimodal information that is obtained through sensors mounted on the robots in an unsupervised manner (Nakamura et al., 2014; Taniguchi et al., 2017).

However, these models are not truly bottom-up because suitable feature extractors for forming concepts are not obtained in a bottom-up manner, but provided manually.

Furthermore, although concept formation is an unsupervised process, the feature extractors include supervised learning, such as a pre-trained convolutional neural network (CNN). These models are based on the MLDA and the relationships among the modality features are learned. The model proposed in (Nakamura et al., 2014) enables robots to acquire object concepts and language simultaneously by mutual learning of the MLDA and language model from multimodal information. The multimodal information is composed of visual, auditory, and tactile information obtained by robots, as well as speech uttered by a human. They are obtained by observing, shaking, and grasping objects, and the human teaches object features through speech. The features are extracted from the visual, auditory, and tactile information by using the dense scale invariant feature transform (Vedaldi and Fulkerson, 2010), sigmoid function approximation, and the Mel-frequency cepstrum coefficient, respectively, which are used for the observation of the model. The model of (Taniguchi et al., 2017), combines the MLDA, simultaneous localization and mapping, the Gaussian mixture model (GMM), and language models to acquire the spatial concept and vocabulary. The human speech to describe the location, visual information of the location, and position of the robot are used to train the model. Similar to the previous method, the features are extracted from each set of information. For example, one of the observations of the model is the feature extracted by the pre-trained CNN known as Places205-AlexNet (Zhou et al., 2014).

In this study, we propose a method that enables unsupervised multimodal clustering, unsupervised feature learning, and cross-modal inference by integrating the MLDA-based concept formation and VAE-based feature extractor. The latent variables in the normal VAE are assumed to follow a Gaussian distribution and these cannot be used for MLDA observations because the observations are assumed to follow multinomial distributions. Jang et al. realized sampling from a categorical distribution using the Gumbel-Softmax distribution (Jang et al., 2017). However, the latent variables were sampled from a categorical distribution and not from a multinomial distribution. Therefore, we extend the VAE to the multinomial VAE (MNVAE), the latent variables of which follow a multinomial distribution, by modifying the method proposed in (Jang et al., 2017). Thereafter, we construct an integrated model of the MNVAE and MLDA and demonstrate that it can classify multimodal information that is composed of images and words. The integrated model performs clustering and learns features simultaneously by communicating the parameters computed in each model. The main contributions of this study are summarized as follows:

- We extend the VAE to the MNVAE, the latent variables of which follow a multinomial distribution, thereby enabling its combination with the MLDA.
- We demonstrate that the clustering performance can be improved and suitable features can be learned by interaction between the MNVAE and MLDA.
- We demonstrate that the cross-modal inference of images from words is possible while learning with a small dataset.

Moreover, we consider that the integration of the MLDA and representation learning is also significant in terms of leveraging past research. Although pre-designed or pre-trained feature extractors have been used, various studies (Attamimi et al., 2014; Nakamura and Nagai, 2017; Hagiwara et al., 2019; Miyazawa et al., 2019) in addition to the above have revealed the effectiveness of the MLDA for multimodal concept formation. That is, the integration of the MLDA and representation learning makes it possible to develop these studies further. Another advantage of the MLDA is that it can be interpreted easily and trained with a small-scale dataset, as indicated in our previous studies. Although it is possible to construct such a model using only deep neural networks (DNNs), it becomes more difficult to interpret the model as its size increases and a larger dataset is required for its training.

In the experiments, we demonstrated that suitable features for clustering images can be learned by the interaction between the MNVAE and MLDA in an unsupervised manner using our integrated model. Furthermore, the integrated model can generate images from the features that are estimated from the words using the MLDA. We used a multimodal dataset composed of images and teaching utterances, which were obtained by the robot observing the objects and the human teacher teaching the object features using speech (Aoki et al., 2016). Because the human teacher did not necessarily utter the words corresponding to the object labels, the teaching utterances included words that were not related to the labels, and speech recognition errors consequently occurred. Moreover, we assumed that the robot did not have a feature extractor for the images in advance. Therefore, the robot was required to learn important words and features in an unsupervised manner from this ambiguous dataset.

2 RELATED WORK

Stochastic models and non-negative matrix factorization have been used in several studies to learn the relationships among multimodal information (Ridge et al., 2010; Nakamura et al., 2014; Mangin et al., 2015; Taniguchi et al., 2017). However, in these works, the feature extractor was designed or learned in advance and it did not learn suitable features using only the training dataset.

Deep generative models such as the VAE and generative adversarial network (Goodfellow et al., 2014) have attracted increasing attention as methods to obtain features. These methods can deal with complicated high-dimensional data in an end-to-end unsupervised manner. In particular, the VAE encoder models the inference process of the latent variables from observations, whereas its decoder models the generative process of the observations from the latent variables. Therefore, the encoder can be used for feature extraction.

The VAE has also been extended to models that can learn the relationships among multimodal information (Suzuki et al., 2017; Wu and Goodman, 2018). In these studies, the features of multimodal information were extracted in an end-to-end manner using multiple encoders and decoders. However, a

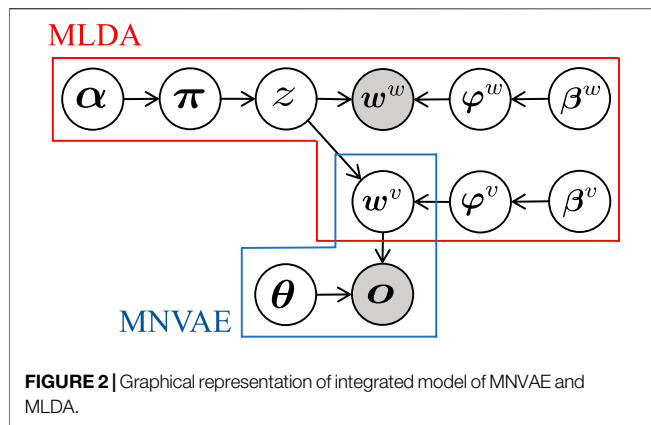


TABLE 1 | Parameters of integrated model.

Parameter	Explanation
\mathbf{o}	Image observation
\mathbf{w}^v	Image feature
ϕ, θ	Parameters of MNVAE encoder and decoder
\mathbf{w}^w	Word observation
z	Category
π	Parameter of multinomial distribution for category
ϕ^v	Parameter of multinomial distribution for image feature
ϕ^w	Parameter of multinomial distribution for word observation
α, β^v, β^w	Parameters of prior distributions

large amount of data was required to train the models, and these were evaluated using a large-scale dataset such as MNIST. It is very difficult to construct a large-scale multimodal dataset using sensors on the robot, and no such dataset exists at present. In our proposed method, the relationships among the multimodal data can be learned from a relatively small-scale multimodal dataset by integrating the MNVAE and MLDA, without designing a feature extractor.

Although the latent variables follow a Gaussian distribution in the normal VAE, models have been proposed in which other distributions were assumed (Jang et al., 2017; Srivastava and Sutton, 2017; Joo et al., 2019). These models made it possible to sample from the distribution, except for a Gaussian distribution, by formulating the sampling procedure into a differentiable form. Jang et al. proposed sampling from a Gumbel-Softmax distribution to obtain samples from a categorical distribution using the Gumbel-Max trick (Gumbel, 1954; Maddison et al., 2014), in which a Gumbel distribution was used instead of the indifferentiable argmax operation (Jang et al., 2017). In the Gumbel-Softmax distribution, the samples became one-hot vectors when the temperature parameter was lower; therefore, the generated samples could be treated as samples from a categorical distribution. Srivastava et al. used the Laplace approximation, whereby the parameters of a Dirichlet distribution were approximated by the parameters of a Gaussian distribution, making it possible to sample from a Dirichlet distribution by adding a softmax layer to the normal

VAE (Srivastava and Sutton, 2017). Joo et al. applied parameter estimation of a multivariate Gamma distribution using an inverse Gamma cumulative distribution function approximation to enable sampling from a Dirichlet distribution (Joo et al., 2019). In this study, we used the Gumbel-Softmax distribution to obtain samples from a multinomial distribution.

Furthermore, clustering methods using features extracted by DNNs have been proposed (Huang et al., 2014; Xie et al., 2016; Yang et al., 2017; Abavisani and Patel, 2018; Hu et al., 2019). Huang et al. added new regularization terms to the feature extraction network to enable the learning of k-means-friendly feature extraction (Huang et al., 2014). Abavisani et al. used DNNs to compute a suitable similarity matrix as an input for spectral clustering (Ng et al., 2002) from multimodal data (Abavisani and Patel, 2018). Although these methods enabled the computation of suitable features for clustering, they did not consider the interaction between clustering and feature extraction. Xie et al. proposed a method to learn the feature extractor and perform clustering simultaneously by fine-tuning the pre-trained autoencoder using the Student's t-distribution (Maaten and Hinton, 2008) and Kullback-Leibler (KL) divergence (Xie et al., 2016). Yang et al. and Hu et al. enabled the simultaneous learning of the feature extractor and clusters based on DNNs and k-means, respectively (Yang et al., 2017; Hu et al., 2019). Hu et al. dealt with the multimodal information of images and audio. Such deep clustering models enable the clustering of complicated data owing to the expressive power of deep neural networks. However, these are deterministic methods that focus only on the task of clustering and cannot perform probabilistic inference, generation, and prediction among multimodal information. Moreover, large datasets are required for training the models. It is difficult to construct such large datasets that are obtained by the robots.

3 INTEGRATED MODEL OF MULTINOMIAL VAE AND MULTIMODAL LATENT DIRICHLET ALLOCATION

Figure 2 presents a graphical illustration of the integrated model of the MNVAE and MLDA. In Figure 2, α , β^v , and β^w are the hyperparameters that determine the Dirichlet priors. Moreover, π , ϕ^v , and ϕ^w are the parameters of the multinomial distribution, whereas θ is a parameter of the MNVAE. \mathbf{o} and \mathbf{w}^w are observations such as an image and a word included in teaching an utterance. \mathbf{w}^v is a feature extracted from \mathbf{o} , and z is a category obtained by clustering \mathbf{w}^v and \mathbf{w}^w . The parameters of the integrated model are listed in Table 1. We aim to extract the image feature \mathbf{w}^v and to form the object category z from the object image \mathbf{o} and word \mathbf{w}^w in an unsupervised manner. In this section, we first explain the components of the integrated model, namely the MNVAE and MLDA, and subsequently describe the parameter estimation of the integrated model. Finally, we clarify the data generation by the MNVAE from a feature predicted by the MLDA.

3.1 Multinomial VAE

In the MNVAE, an observation \mathbf{o} is compressed into an arbitrary-dimensional latent variable \mathbf{w}^v through a neural network known as an encoder. Thereafter, \mathbf{w}^v is reconstructed into the original dimensional value $\hat{\mathbf{o}}$ through a neural network known as a decoder. In this case, ϕ and θ denote the parameters of the encoder and decoder, respectively. The parameters are learned such that $\hat{\mathbf{o}}$ becomes the same as \mathbf{o} . However, in the normal VAE, it is assumed that a prior follows a Gaussian distribution, and therefore it cannot share a latent variable with the MLDA, in which it is assumed that the observations follow multinomial distributions. To address this problem, the MNVAE computes λ_k , which are the parameters of the multinomial distribution $q_\phi(\mathbf{w}^v|\mathbf{o})$, where $k = 1, 2, \dots, K$, and K is the number of dimensions. Subsequently, sampling from the multinomial distribution is performed according to the following operation:

$$u_{n,k} \sim \text{Uniform}(0, 1), \quad (1)$$

$$g_{n,k} = -\log(-\log(u_{n,k})), \quad (2)$$

$$w_{n,k}^v = \frac{\exp((\log(\lambda_k) + g_{n,k})/\tau)}{\sum_{k=1}^K \exp((\log(\lambda_k) + g_{n,k})/\tau)}. \quad (3)$$

$$\mathbf{w}^v = \sum_{n=1}^N \mathbf{w}_n^v, \quad (4)$$

where N is the number of samplings and τ is the temperature. By setting τ to a small value, \mathbf{w}_n^v becomes a one-hot vector.

3.2 Multimodal Latent Dirichlet Allocation

The MLDA is a model that extends the LDA (Blei et al., 2003) and can classify multimodal information. In the MLDA, it is assumed that the observations $\mathbf{w}^v = \{w_1^v, \dots, w_{N^v}^v\}$ and $\mathbf{w}^w = \{w_1^w, \dots, w_{N^w}^w\}$ are generated as follows, in which N^v and N^w are the total numbers of features for each observation.

- The category proportion π is determined by the Dirichlet prior $P(\pi|\alpha)$:

$$\pi \sim P(\pi|\alpha). \quad (5)$$

- The following procedure is repeated N^m times for $m \in \{v, w\}$ to generate the observations.

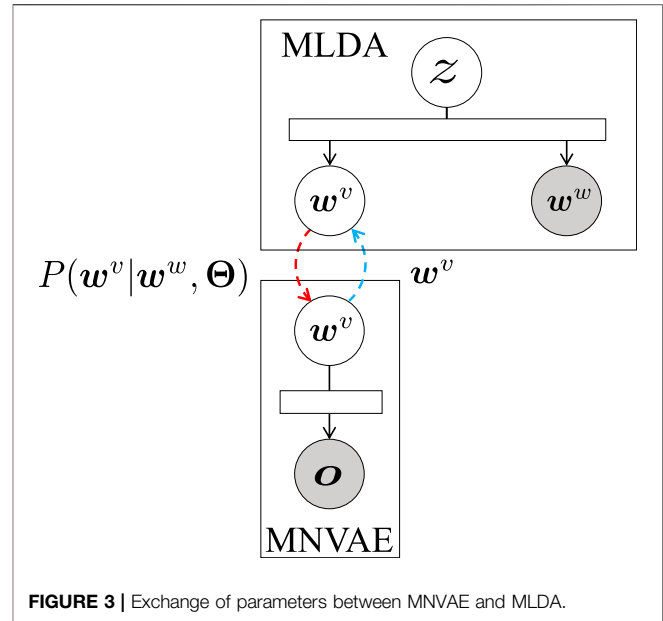
- A category z is sampled from the multinomial distribution $P(z|\pi)$:

$$z \sim P(z|\pi). \quad (6)$$

- The observation w_{nm}^m corresponding to the category z is generated from the multinomial distribution $P(w^m|\phi_z^m)$:

$$w_{nm}^m \sim P(w^m|\phi_z^m). \quad (7)$$

The observations \mathbf{w}^v and \mathbf{w}^w are classified by estimating the category z and parameters π , ϕ^v , and ϕ^w using Gibbs sampling in an unsupervised manner. Sampling is conducted



such that the category z^{mij} is assigned to the i th feature of the modality $m \in \{v, w\}$ information \mathbf{w}^m of the j th object from the conditional probability, where π and ϕ^m are marginalized out as follows:

$$P(z^{mij} = c | z^{-mij}, \mathbf{w}^m, \alpha, \beta^m) \propto (N_{cj}^{-mij} + \alpha_c) \frac{N_{mw_i^m c}^{-mij} + \beta_i^m}{N_{mc}^{-mij} + D^m \beta_i^m}, \quad (8)$$

in which D^m is the dimension number of the observation of modality m . Moreover, a negative superscript indicates that the information is excluded. For example, z^{-mij} represents the remainder of the set of categories assigned to all objects, excluding z^{mij} . $N_{mw_i^m c}$ is the number of times that category c is assigned to w_i^m , which is the i th feature of modality m of the j th object. Furthermore, $N_{mw_i^m c}$, N_{cj} , N_{mc} are computed as follows:

$$N_{mw_i^m c} = \sum_j N_{mw_i^m c j}, \quad (9)$$

$$N_{cj} = \sum_{m, w_i^m} N_{mw_i^m c j}, \quad (10)$$

$$N_{mc} = \sum_{w_i^m} N_{mw_i^m c j}. \quad (11)$$

That is, $N_{mw_i^m c}$ is the number of times that category c is assigned to the i th feature w_i^m of modality m of all objects, N_{cj} is the number of times that category c is assigned to all modalities of the j th object, and N_{mc} is the number of times that category c is assigned to the observation \mathbf{w}^m of the modality m of all objects. By means of repeated sampling according to Eq. 8, N^* converges to N^* and the parameters are computed as follows:

$$\hat{\phi}_{mw_i^m c} = \frac{\bar{N}_{mw_i^m c} + \beta_i^m}{\bar{N}_{mc} + D^m \beta_i^m}, \quad (12)$$

$$\hat{\pi}_{cj} = \frac{\bar{N}_{cj} + \alpha_c}{N_j + C\alpha_c}, \quad (13)$$

where N_j is the total number of features of all modalities of the j th object and C is the number of categories. Finally, category z_j of the j th object can be determined as follows:

$$z_j = \operatorname{argmax}_c \hat{\pi}_{cj}. \quad (14)$$

3.3 Parameter Estimation of Integrated Model

In this section, we discuss the parameter estimation of the integrated model. It is desirable for the model parameters illustrated in **Figure 2** to be optimized by directly maximizing the likelihood. However, it is difficult to estimate the latent variable \mathbf{w}^v because it is shared between the MNVAE and MLDA. Therefore, in this study, the MNVAE and MLDA are learned independently, and all parameters are optimized approximately by exchanging the parameters computed in each model.

Figure 3 presents the exchange of parameters between the MNVAE and MLDA. The MNVAE compresses the observation \mathbf{o} into \mathbf{w}^v through the encoder and sends it to the MLDA. The MLDA considers \mathbf{w}^v as an observation, and classifies \mathbf{w}^v and observation \mathbf{w}^w into category z . Thereafter, it sends the parameters of the distribution $p(\mathbf{w}^v|\mathbf{w}^w, \Theta)$ to the MNVAE, where Θ is a set of parameters learned in the MLDA. Subsequently, the parameters of the MNVAE are optimized by the variational lower bound using the received parameters, as follows:

$$\begin{aligned} \mathcal{L}(\theta, \phi; \mathbf{o}) = & -\gamma D_{KL}(q_\phi(\mathbf{w}^v|\mathbf{o})\|P(\mathbf{w}^v|\mathbf{w}^w, \Theta)) \\ & + \mathbb{E}_{q_\phi(\mathbf{w}^v|\mathbf{o})}[\log p_\theta(\mathbf{o}|\mathbf{w}^v)], \end{aligned} \quad (15)$$

where D_{KL} represents the KL divergence and γ is the weight thereon. In the integrated model, the MNVAE is initially optimized by using a uniform distribution because the order of the parameter update is MNVAE \rightarrow MLDA \rightarrow MNVAE $\rightarrow \dots$, and $p(\mathbf{w}^v|\mathbf{w}^w, \Theta)$ has not yet been computed at this time. Thus, a latent space that is suitable for clustering is learned by regularizing $q_\phi(\mathbf{w}^v|\mathbf{o})$, using $p(\mathbf{w}^v|\mathbf{w}^w, \Theta)$ as a prior.

We use the Symbol Emergence in Robotics tool KIT (SERKET) (Nakamura et al., 2018; Taniguchi et al., 2020) to implement the integrated model. SERKET is a framework for constructing a large-scale model composed of small models. SERKET supports independent learning in each module, the exchange of computed parameters between modules, and the optimization of the parameters of an entire model by means of the interaction among modules, as described above.

3.4 Cross-Modal Inference Using Integrated Model

It is possible to predict an observation from another observation using the learned parameters in the integrated model. When a new observation $\mathbf{w}^{w'}$ is obtained, the probability $p(\mathbf{w}^v|\mathbf{w}^{w'}, \Theta)$ that the unobserved image feature \mathbf{w}^v will be generated from the words $\mathbf{w}^{w'}$ can be computed. The image feature $\widehat{\mathbf{w}}^v$ is constructed by performing the following sampling from this distribution N times:

$$\mathbf{w}^v \sim P(\mathbf{w}^v|\mathbf{w}^{w'}, \Theta), \quad (16)$$

$$\widehat{\mathbf{w}}^v[\mathbf{w}^v] += 1. \quad (17)$$

The constructed $\widehat{\mathbf{w}}^v$ is input into the MNVAE decoder $p_\theta(\mathbf{o}|\widehat{\mathbf{w}}^v)$, and an image \mathbf{o}' that is predicted from $\mathbf{w}^{w'}$ can subsequently be generated.

$$\mathbf{o}' \sim p_\theta(\mathbf{o}|\widehat{\mathbf{w}}^v). \quad (18)$$

4 EXPERIMENTS

4.1 Dataset

In the experiments, we used the dataset that was used in (Aoki et al., 2016). This dataset was composed of images, tactile sensor values, and sound data obtained by the robot from objects and utterances given by a human who teaches features of the objects. The number of objects used was 499, which included a total of 81 categories, such as plastic bottles, candy boxes, stuffed animals, and rattles. We used only the image and utterance data from the dataset in the experiments. The images were obtained by extracting the object region using object detection from the scene acquired by the RGB camera. The images were resized to 144×120 because the image size differed according to the object size. We used these images as the observation \mathbf{o} . The teaching utterances were converted into strings using a phoneme recognizer and then divided into words in an unsupervised manner using the pseudo-online NPYLM (Araki et al., 2013). These words were converted into bag of words (BoWs) representations, and we used these BoWs as the observation \mathbf{w}^w .

4.2 Clustering and Representation Learning

We classified the multimodal dataset composed of images \mathbf{o} and words \mathbf{w}^w using the integrated model. We iterated the training of the integrated model by varying the initial value 10 times, and used the average result to evaluate the proposed method. **Figure 4** presents the network architecture of the MNVAE. We set $K = 32$, $N = 100$, and $\gamma = 1$ in the first learning and $\gamma = 10^4$ thereafter. Moreover, learning was performed with a batch size of 100 and 3,000 epochs. In this case, the initial value of τ was set to 1, and it was decayed at a rate of $\exp(-10^{-3}t)$ for each epoch t . The MLDA was optimized by Gibbs sampling with 100 sampling times.

As a comparison method, the features extracted by the joint multimodal VAE (JMVAE) (Suzuki et al., 2017) were classified by the GMM. In the JMVAE, the encoder and decoder for the images have the same structure as those of the MNVAE; that is, the encoder has three convolutional layers and fully connected (FC) layers, whereas the decoder has an FC layer and seven deconvolutional layers. The number of dimensions of the latent variables was set to 32, as in the MNVAE. Refer to the **Appendix 1** for further details on the structure.

Figure 5 presents the latent variables compressed into two dimensions by t-SNE (Maaten and Hinton, 2008) for visualization. Each point represents one object and the color reflects its correct category. **Table 2** displays the classification

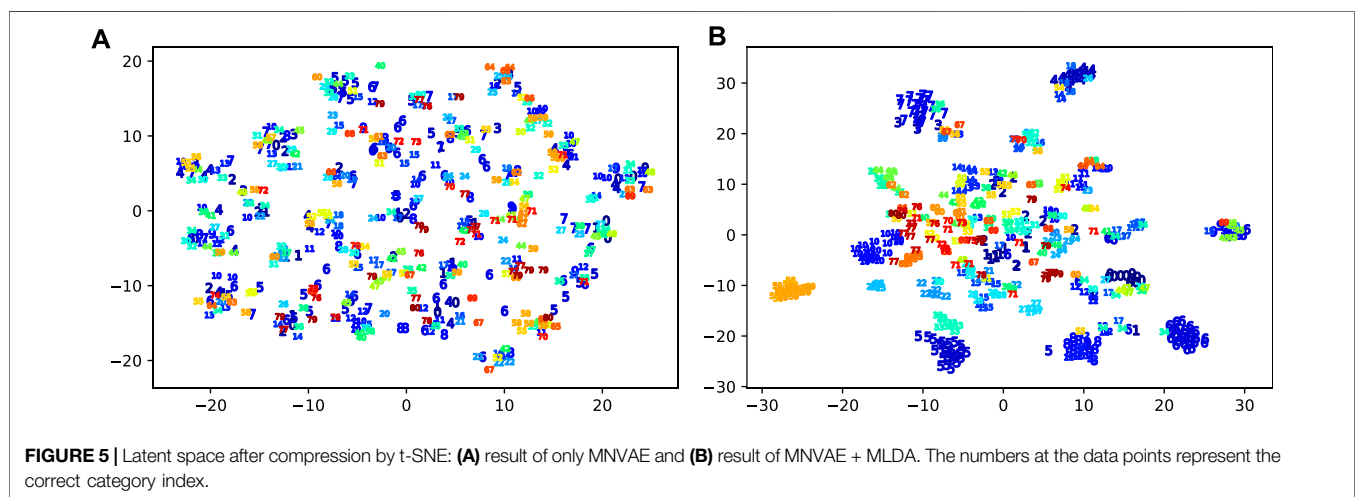
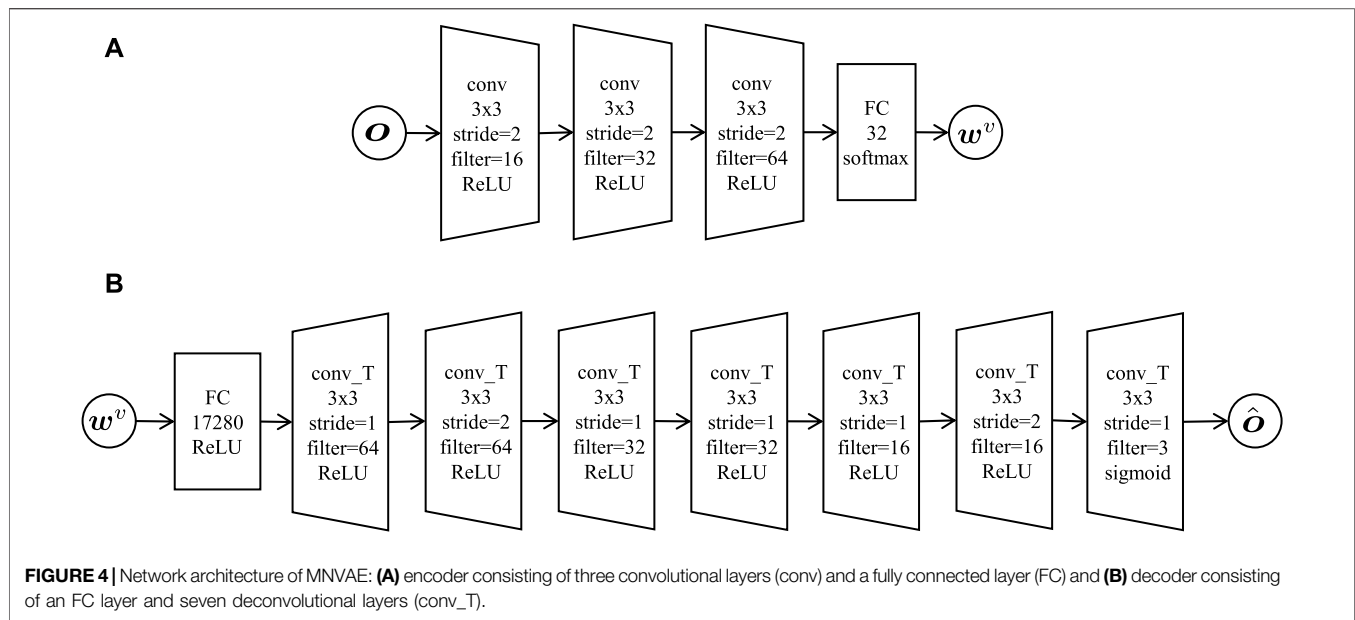


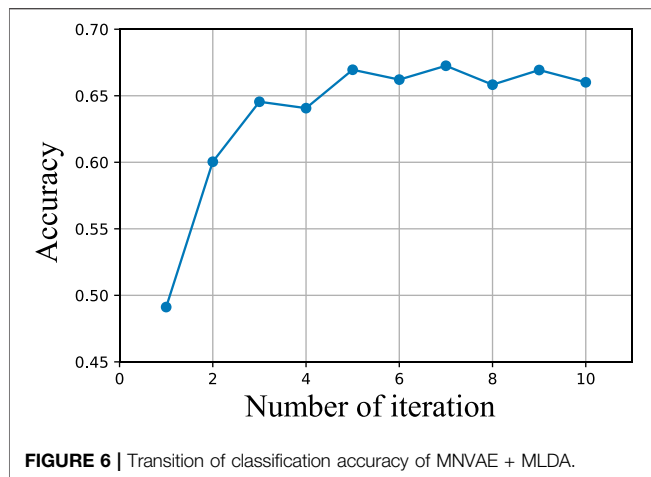
TABLE 2 | Classification accuracies and ARIs. “MLDA” used a pre-trained CNN for the feature extraction, and is therefore considered as the performance upper bound.

	Accuracy	ARI
MLDA	0.688 ± 0.014	0.613 ± 0.026
JMVAE + GMM	0.465 ± 0.013	0.271 ± 0.016
MNVAE-MLDA	0.491 ± 0.024	0.374 ± 0.031
MNVAE + MLDA	0.660 ± 0.018	0.601 ± 0.032

accuracies and adjusted rand indices (ARIs) (Hubert and Arabie, 1985). “MLDA” is the result of the MLDA classifying the words and image features extracted using the pre-trained CNN model¹

¹https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet

(Krizhevsky et al., 2012; Jia et al., 2014), as in the previous study (Aoki et al., 2016); “JMVAE + GMM” is the result of the GMM classifying the features extracted by the JMVAE; “MNVAE-MLDA” is the result of the classification by the integrated model without interaction; and “MNVAE + MLDA” is the result of the classification by the integrated model with 10 interactions. Moreover, the transition of the classification accuracy of the MNVAE + MLDA by exchanging the parameters is depicted in **Figure 6**. The horizontal axis represents the number of exchange iterations and the vertical axis indicates the accuracy. **Figure 5A** indicates that the data points of the same correct category were not close to one another when using only the MNVAE. However, they were located close together with the interaction between the MNVAE and MLDA, as illustrated in **Figure 5B**. In particular, the data points of categories 5, 6, and 59 were relatively well separated in the latent space. This was because the dataset exhibited a bias



in the number of objects included in each category, and the numbers of objects in categories 5, 6, and 59 were relatively large compared to those of the other categories. Furthermore, **Figure 6** indicates that the classification accuracy was improved by iterating the exchange of

parameters between the MNVAE and MLDA. Thus, the object category and word information affected the MNVAE through the interaction, and feature extraction that was suitable for clustering was obtained, thereby improving the classification accuracy. **Table 2** indicates that the classification accuracy of the unsupervised MNVAE + MLDA was similar to that of the MLDA with the supervised feature extraction and better than that of the JMVAE + GMM, which is likewise a method for unsupervised feature extraction and clustering. This suggests the effectiveness of the proposed learning method, in which the MNVAE and MLDA parameters are exchanged.

4.3 Cross-Modal Inference From Words to Images

This experiment generated images from words, and the generated images were evaluated to show that the proposed method can perform cross-modal inference. The moderate quality of the generated images may be attributed to the fact that the training dataset was not sufficiently large. However, it is considered that MNVAE + MLDA can

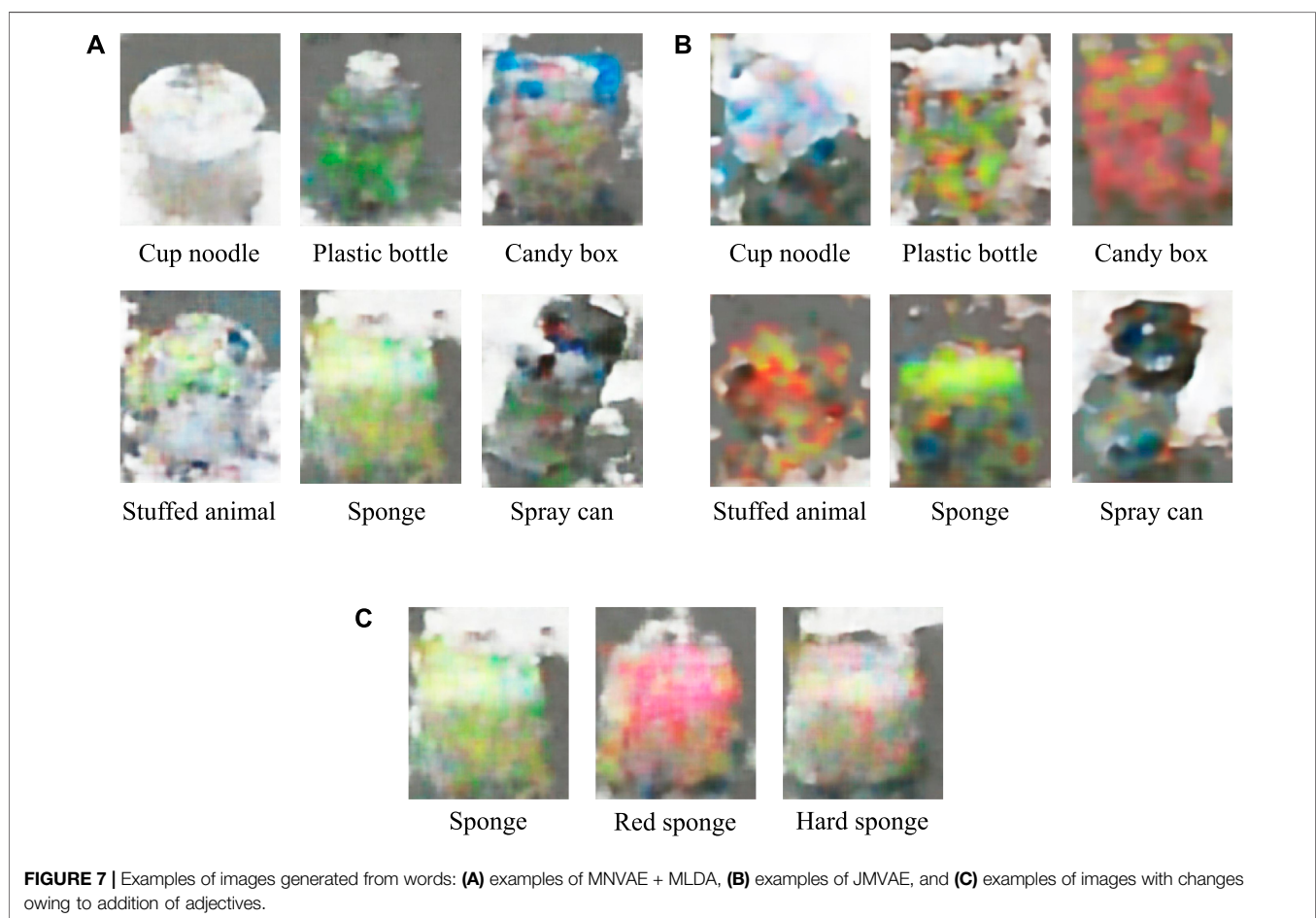


TABLE 3 | Results of subjective evaluation of generated images. This table indicates the number of subjects who selected each method as generating more appropriate images representing the word.

Word	MNVAE + MLDA	JMVAE
Cup noodle	18	0
Plastic bottle	17	1
Candy box	18	0
Stuffed animal	14	4
Sponge	8	10
Spray can	17	1

generate more word-representative images than JMVAE because a latent space that is suitable for clustering is learned by regularizing $q_{\phi}(\mathbf{w}''|\mathbf{o})$ using $p(\mathbf{w}''|\mathbf{w}', \Theta)$ in Eq. 15. Therefore, we conducted a subjective experiment to identify the method that can generate more word-representative images.

Images were generated from words using the MNVAE + MLDA learned in the clustering experiment. The words included “cup noodle”, “plastic bottle”, and “candy box”. **Figure 7A,B** present examples of the images generated from the words using the trained MNVAE + MLDA and JMVAE. To evaluate the generated images, we showed 18 subjects a word and two images generated from the word by the MNVAE + MLDA and JMVAE, and asked them to choose which image was more appropriate to represent the word. This procedure was iterated for six words. The number of subjects who selected each method is displayed in **Table 3**. As illustrated in **Figure 7A**, the rough shapes of the objects were produced, although they were blurry. A white cylinder object was produced for “cup noodle” and a green bottle with a white cap was produced for “plastic bottle”. The dataset images included numerous white cup noodles, green plastic bottles, and biscuits and other sweets contained in boxes, as well as yellow and blue sponges. Therefore, the images capturing these features were generated. In contrast, many of the images in **Figure 7B** were very collapsed and it was difficult to recognize them as the objects. This is because the dataset used in this experiment has only 499 data points, which is relatively small for training the JMVAE. Moreover, **Table 3** demonstrates that the MNVAE + MLDA-generated images were selected as the more word-representative images for most objects in the subjective experiment. Although the image generated from the sponge was comparable, this is because the sponge is a simple shaped object that the JMVAE could generate. According to these results, the MNVAE + MLDA was better able to perform cross-modal inference in the environment of this experiment.

Furthermore, it was confirmed that the generated images were changed by the addition of adjectives. Examples of this change are presented in **Figure 7C**. An image of a red sponge with a slight yellow tint was generated by adding “red” to “sponge” and “hard sponge” generated an image of a grayish sponge. It is believed that this was because “hard” was often taught for gray cans. Thus, the MNVAE + MLDA could learn suitable features from the small dataset, making it possible not only to compute the probability of generating the feature values, but also to generate the actual images.

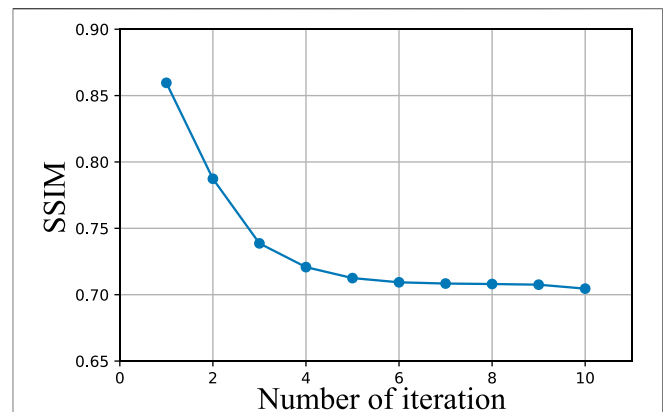


FIGURE 8 | Transition of SSIM.

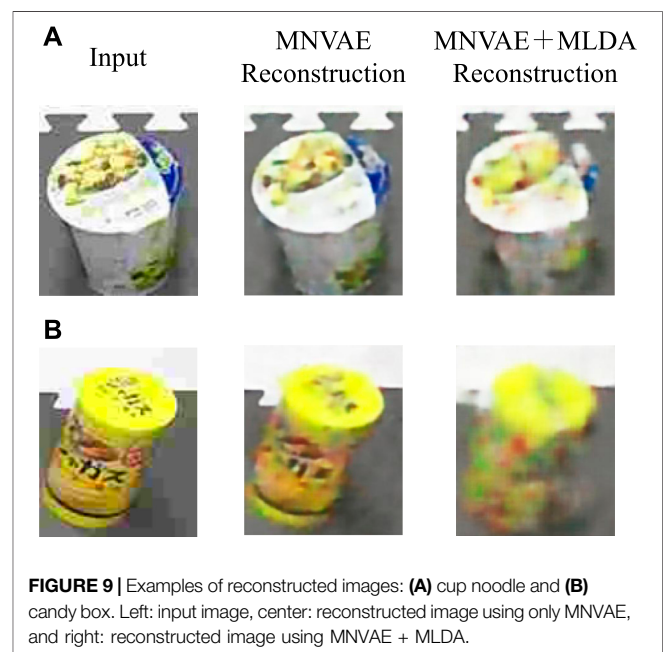


FIGURE 9 | Examples of reconstructed images: **(A)** cup noodle and **(B)** candy box. Left: input image, center: reconstructed image using only MNVAE, and right: reconstructed image using MNVAE + MLDA.

5 DISCUSSION

5.1 Image Reconstruction

We computed the structural similarity (SSIM) (Wang et al., 2004) as a quantitative evaluation of the reconstructed images. The transition of the SSIM of the MNVAE + MLDA by exchanging the parameters is illustrated in **Figure 8**. The horizontal axis represents the number of exchange iterations and the vertical axis indicates the SSIMs. Moreover, **Figure 9** presents examples of images reconstructed by the MNVAE. It can be observed from **Figure 8** that the SSIM of the MNVAE + MLDA decreased with the interactions. This is because the latent variables represent not only the information for reconstructing the images, but also the features of the categories owing to the parameters received from the MLDA. Therefore, the SSIM

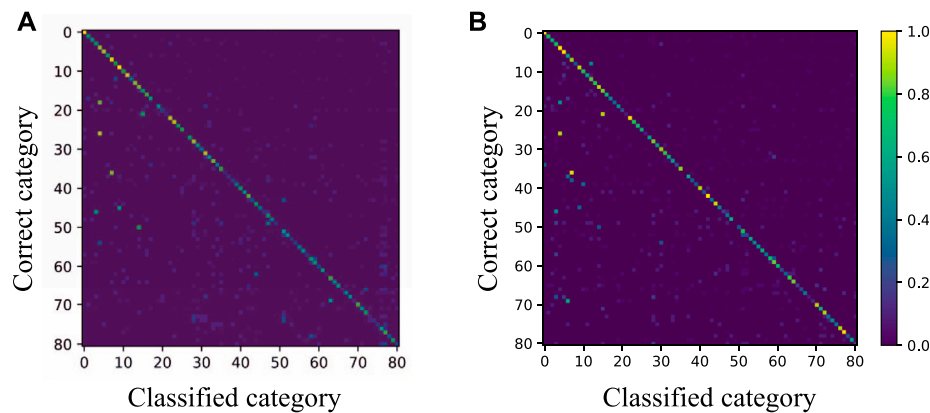


FIGURE 10 | Classification results (normalized confusion matrices): **(A)** MNVAE + MLDA and **(B)** MLDA.

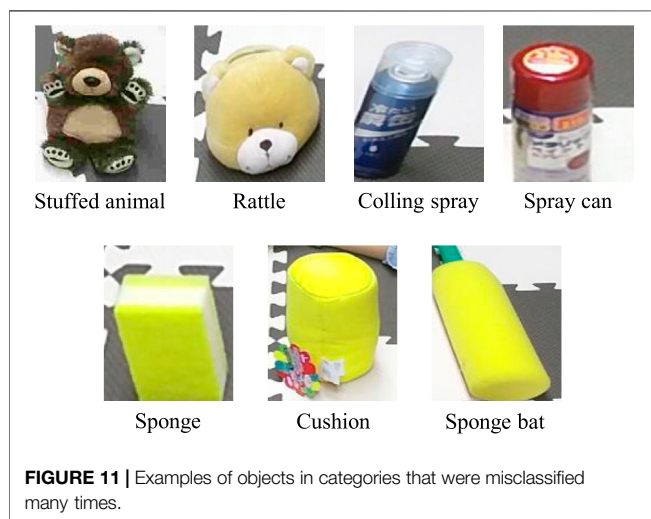


FIGURE 11 | Examples of objects in categories that were misclassified many times.

decreased and the reconstruction error increased. Although the SSIM was decreased, it was possible to distinguish objects from the reconstructed images, as indicated in **Figure 9**, suggesting that feature extraction that was suitable for clustering could be performed while capturing the image features for their reconstruction.

5.2 Classification Results

Figure 10 presents the classification results of the MNVAE + MLDA and MLDA. **Figures 10A,B** are normalized confusion matrices. The vertical axis is the index of the correct category and the horizontal axis is the index of the classified category. **Figure 10** indicates that stuffed animals (category 7) and rattles (category 36); cooling sprays (category 21) and spray cans (category 15); and sponges (category 4), cushions (Category 18), and sponge bats (category 26) were frequently misclassified in the MNVAE + MLDA and MLDA. Examples of the objects included in these categories are depicted in **Figure 11**. As illustrated in **Figure 11**, the features of these objects were quite

similar. Furthermore, the features of the rattles were often taught by using utterances such as “this is a stuffed animal that makes sound”, whereas the features of the sponge, cushion, and sponge bat were taught by using utterances such as “this is soft”. Therefore, these categories had similar image features and similar words such as “stuffed animal”, “spray”, and “soft” were taught frequently, thereby increasing the misclassification. In (Aoki et al., 2016), tactile and sound information was used in addition to images and words, and the correct categories were determined on this basis. Hence, it is possible to decrease the misclassification by adding information that was not used in the experiment. For example, as the sounds of stuffed animals and rattles differ when shaking them, it is possible to classify them correctly using sound information. Comparing **Figures 10A,B**, it can be observed that the MNVAE + MLDA yielded slightly more misclassifications than the MLDA. The MLDA used the features extracted by the pre-trained CNN, which was effective. In contrast, in the MNVAE + MLDA, the feature extractor was learned at the same time by the interaction between the MNVAE and MLDA. As a result, we consider that several objects were misclassified because of biased words, and the features were also learned to represent those misclassifications.

5.3 Definition of Concepts

This paper uses a straightforward definition of concepts such as the multimodal categories that enable multimodal clustering and cross-modal inference, to present our outcome. Moreover, we deal with categories such as discretized and symbolic to compare them with the manually determined ground truth. This is because we focus on evaluating our proposed method from the perspective of engineering. By contrast, in cognitive science (CS), the characteristics of concepts have been studied (Olier et al., 2017), and we consider that our MLDA-based approach has the potential to replicate some of them.

One characteristic pointed out in CS is that a concept is not symbolic, but a distributed representation. In our approach, although we convert categories into the symbolic representation z by $\arg \max_p (z|w', w'')$, the categories can be

represented by the distribution $p(z|w', w'')$, which has a continuous parameter. Therefore, concepts are represented in a continuous space in MLDA, and thus it does not represent only symbolic and static categories.

It is also pointed out that embodiment is important and that concepts depend on action and context. An action can be easily introduced into MLDA by considering the action a to be a single modality as $p(z|w', w'', a)$ (Fadlil et al., 2013). By connecting perception and action, it is possible to recall the unobserved perception from the action through a simulation $p(w''|a)$. Furthermore, we extend our MLDA to a time-series model, and action- and context-dependent concepts can be learned (Miyazawa et al., 2019) to a certain extent.

Although the MLDA-based approach cannot currently describe all cognitive phenomena, we consider MLDA to be a promising model. In the future, by integrating our previous studies (Fadlil et al., 2013; Miyazawa et al., 2019) and the proposals in this paper, we would like to develop a cognitive model that allows robots to learn by interacting with the environment.

6 CONCLUSIONS

We have proposed the MNVAE, which is an extension of the VAE, the latent variables of which follow a multinomial distribution. An integrated model of the MNVAE and MLDA was constructed, and the multimodal information was classified. The experiments demonstrated that the interaction between the MNVAE and MLDA could learn the features that are suitable for clustering. Moreover, the images representing the concepts acquired by the MLDA can be generated from the word information. Thus, we have revealed that combining the MNVAE and MLDA enables clustering and the learning of features from the information obtained from the environment in a truly bottom-up and unsupervised manner, without the need for a manually designed feature extractor, as used in previous studies.

We evaluated the integrated model using only image and word information. In the future, we will incorporate not only images and words, but also other modal information, such as tactile and auditory information. We will construct and evaluate a model that learns the feature extraction of the information of each modality using the MNVAE in an unsupervised manner. Furthermore, in this study, the BoW representation of strings recognized by a phoneme recognizer was used, and language knowledge (the language model) in the speech recognizer was not updated. We believe that acquiring such language knowledge from the information obtained by its own sensors in a bottom-up manner is important (Tangiuchi et al., 2019). Therefore, we will introduce mutual learning with the language model (Nakamura

et al., 2014) into the integrated model, and we will construct a model that can learn the parameters of the speech recognizer simultaneously. Moreover, the experimental results showed that it was possible to learn from a small amount of data, but we found that certain categories could not be learned correctly. We consider that this is because the number of objects is small for learning these categories and the bias of the number of objects in each category is inevitable in the real environment. To achieve learning in such a biased environment, we believe that interactive and online learning is very important. The learner expresses its inner state by using its current knowledge and body, and the teacher changes his/her actions (e.g., the object presented next in the task used in this paper) by estimating the extent to which the learner could acquire knowledge from its feedback. Therefore, we plan to extend the proposed method to online learning to realize interactive learning.

DATA AVAILABILITY STATEMENT

The datasets used for this study can be found at our GitHub repository <https://github.com/naka-lab/MLDA-MNVAE>.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

RK, TNak, TT, and TNag conceived, designed, and developed the research. RK and TNak performed the experiment and analyzed the data. RK wrote the manuscript with support from TNak, TT, and TNag. TNag supervised the project. All authors discussed the results and contributed to the final manuscript.

FUNDING

This work was supported by JST, Moonshot R and D Grant Number JPMJMS 2011.

REFERENCES

- Abavisani, M., and Patel, V. M. (2018). Deep Multimodal Subspace Clustering Networks. *IEEE J. Sel. Top. Signal. Process.* 12, 1601–1614. doi:10.1109/jstsp.2018.2875385
- Aoki, T., Nishihara, J., Nakamura, T., and Nagai, T. (2016). "Online Joint Learning of Object Concepts and Language Model Using Multimodal Hierarchical Dirichlet Process," in IEEE/RSJ International Conference on Intelligent Robots and Systems, Daejeon, Korea, 2636–2642. doi:10.1109/iro.2016.7759410
- Araki, T., Nakamura, T., and Nagai, T. (2013). "Long-Term Learning of Concept and Word by Robots: Interactive Learning Framework and Preliminary Results," in IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 2280–2287. doi:10.1109/iro.2013.6696675

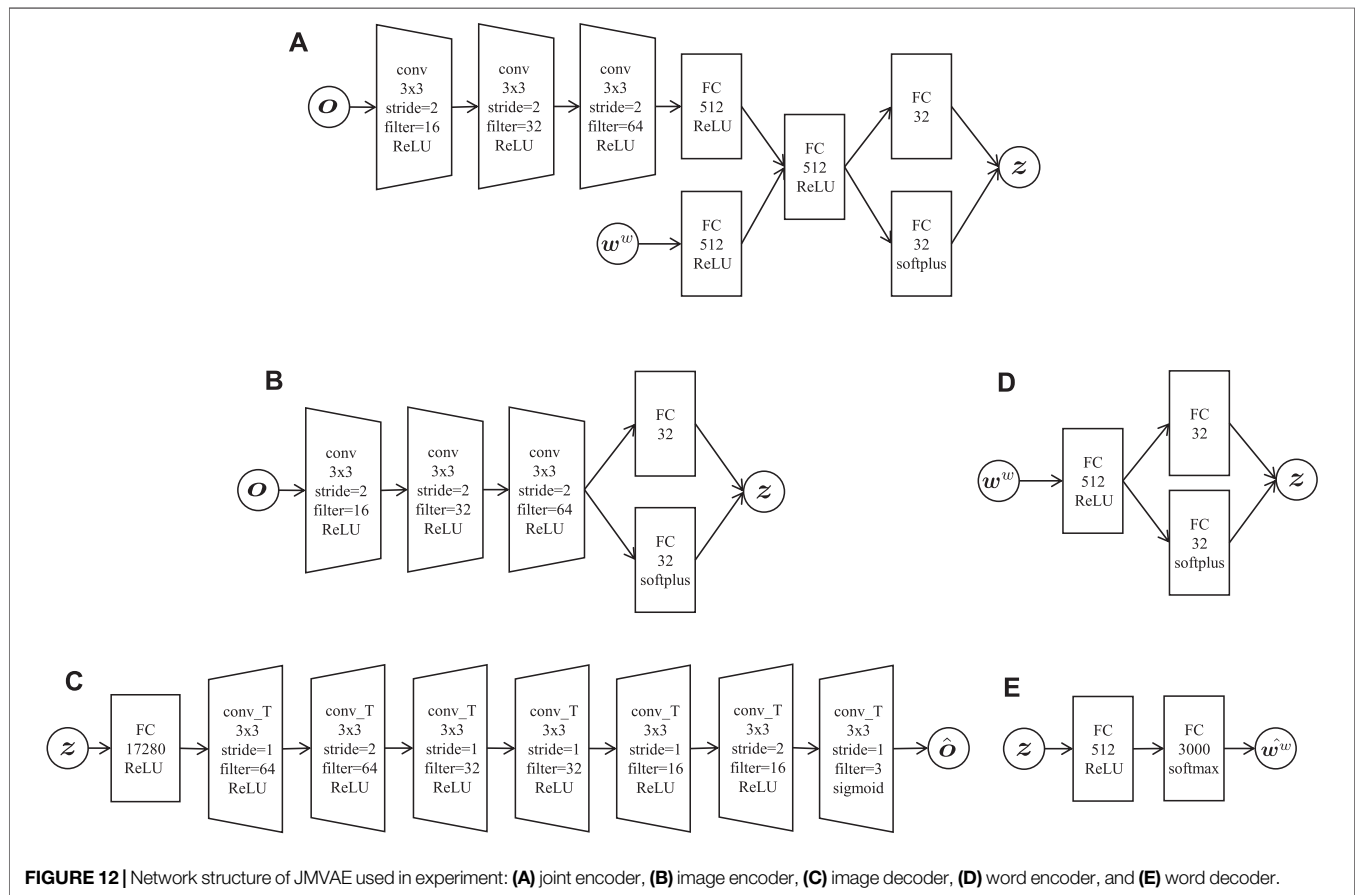
- Attamimi, M., Fadlil, M., Abe, K., Nakamura, T., Funakoshi, K., and Nagai, T. (2014). "Integration of Various Concepts and Grounding of Word Meanings Using Multi-Layered Multimodal Lda for Sentence Generation," in IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 2194–2201. doi:10.1109/iros.2014.6942858
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Machine Learn. Res.* 3, 993–1022.
- Fadlil, M., Ikeda, K.-i., Abe, K., Nakamura, T., and Nagai, T. (2013). "Integrated Concept of Objects and Human Motions Based on Multi-Layered Multimodal Lda," in IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 2256–2263. doi:10.1109/iros.2013.6696672
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative Adversarial Nets," in Advances in neural information processing systems, Montreal, Canada, 2672–2680.
- Gumbel, E. J. (1954). Statistical Theory of Extreme Values and Some Practical Applications. *NBS Appl. Mathematics Ser.* 33, 1–51.
- Hagiwara, Y., Kobayashi, H., Taniguchi, A., and Taniguchi, T. (2019). Symbol Emergence as an Interpersonal Multimodal Categorization. *Front. Robot. AI* 6, 134. doi:10.3389/frobt.2019.00134
- Hu, D., Nie, F., and Li, X. (2019). "Deep Multimodal Clustering for Unsupervised Audiovisual Learning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, 9248–9257. doi:10.1109/cvpr.2019.00947
- Huang, P., Huang, Y., Wang, W., and Wang, L. (2014). "Deep Embedding Network for Clustering," in 2014 22nd International conference on pattern recognition, Stockholm, Sweden, (IEEE), 1532–1537. doi:10.1109/icpr.2014.272
- Hubert, L., and Arabie, P. (1985). Comparing Partitions. *J. Classification.* 2, 193–218. doi:10.1007/bf01908075
- Jang, E., Gu, S., and Poole, B. (2017). "Categorical Reparameterization With Gumbel-Softmax," in 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017 (Conference Track Proceedings).
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. (2014). "Caffe: Convolutional Architecture for Fast Feature Embedding," in ACM International Conference on Multimedia, Orlando, FL, 675–678.
- Joo, W., Lee, W., Park, S., and Moon, I.-C. (2019). Dirichlet Variational Autoencoder. arXiv preprint arXiv:1901.02739, 1–17.
- Kingma, D. P., and Welling, M. (2013). Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114, 1–14.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet Classification With Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 25, 1097–1105. doi:10.1145/3065386
- Maaten, L. v. d., and Hinton, G. (2008). Visualizing Data Using T-Sne. *J. Machine Learn. Res.* 9, 2579–2605.
- Maddison, C. J., Tarlow, D., and Minka, T. (2014). A Sampling. *Adv. Neural Inf. Process. Syst.* 27, 3086–3094.
- Mangin, O., Filliat, D., Ten Bosch, L., and Oudeyer, P.-Y. (2015). Mca-nmf: Multimodal Concept Acquisition With Non-Negative Matrix Factorization. *PLoS one* 10, e0140732. doi:10.1371/journal.pone.0140732
- Miyazawa, K., Horii, T., Aoki, T., and Nagai, T. (2019). Integrated Cognitive Architecture for Robot Learning of Action and Language. *Front. Robot. AI* 6, 131. doi:10.3389/frobt.2019.00131
- Mochihashi, D., Yamada, T., and Ueda, N. (2009). "Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling," in Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing, Suntec, Singapore, 100–108. doi:10.3115/1687878.1687894
- Nagano, M., Nakamura, T., Nagai, T., Mochihashi, D., Kobayashi, I., and Takano, W. (2019). Hvgh: Unsupervised Segmentation for High-Dimensional Time Series Using Deep Neural Compression and Statistical Generative Model. *Front. Robot. AI* 6, 115. doi:10.3389/frobt.2019.00115
- Nakamura, T., Nagai, T., and Taniguchi, T. (2018). Serket: An Architecture for Connecting Stochastic Models to Realize a Large-Scale Cognitive Model. *Front. Neurobot.* 12, 25–16. doi:10.3389/fnbot.2018.00025
- Nakamura, T., and Nagai, T. (2017). Ensemble-of-Concept Models for Unsupervised Formation of Multiple Categories. *IEEE Trans. Cogn. Developmental Syst.* 10, 1043–1057. doi:10.1109/TCDS.2017.2745502
- Nakamura, T., Nagai, T., Funakoshi, K., Nagasaka, S., Taniguchi, T., and Iwahashi, N. (2014). "Mutual Learning of an Object Concept and Language Model Based on Mlda and Npylm," in IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 600–607. doi:10.1109/iros.2014.6942621
- Nakamura, T., Nagai, T., and Iwahashi, N. (2007). "Multimodal Object Categorization by a Robot," in IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, CA, USA, 2415–2420. doi:10.1109/iros.2007.4399634
- Nakamura, T., Nagai, T., and Iwahashi, N. (2009). "Grounding of Word Meanings in Multimodal Concepts Using LDA," in IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, USA, 3943–3948. doi:10.1109/iros.2009.5354736
- Neubig, G., Mimura, M., Mori, S., and Kawahara, T. (2012). Bayesian Learning of a Language Model From Continuous Speech. *IEICE Trans. Inf. Syst.* E95-D, 614–625. doi:10.1587/transinf.e95.d.614
- Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On Spectral Clustering: Analysis and an Algorithm. *Adv. Neural Inf. Process. Syst.* 14, 849–856.
- Olier, J. S., Barakova, E., Regazzoni, C., and Rauterberg, M. (2017). Re-Framing the Characteristics of Concepts and Their Relation to Learning and Cognition in Artificial Agents. *Cogn. Syst. Res.* 44, 50–68. doi:10.1016/j.cogsys.2017.03.005
- Piaget, J., and Duckworth, E. (1970). Genetic Epistemology. *Am. Behav. Scientist.* 13, 459–480. doi:10.1177/000276427001300320
- Ridge, B., Skocaj, D., and Leonardis, A. (2010). "Self-Supervised Cross-Modal Online Learning of Basic Object Affordances for Developmental Robotic Systems," in IEEE International Conference on Robotics and Automation, Anchorage, AK, USA, 5047–5054. doi:10.1109/robot.2010.5509544
- Srivastava, A., and Sutton, C. A. (2017). "Autoencoding Variational Inference for Topic Models," in 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings.
- Suzuki, M., Nakayama, K., and Matsuo, Y. (2017). "Joint Multimodal Learning With Deep Generative Models," in 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Workshop Track Proceedings (OpenReview.net).
- Taniguchi, A., Hagiwara, Y., Taniguchi, T., and Inamura, T. (2017). Online Spatial Concept and Lexical Acquisition With Simultaneous Localization and Mapping. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada. doi:10.1109/iros.2017.8202243
- Taniguchi, T., Mochihashi, D., Nagai, T., Uchida, S., Inoue, N., Kobayashi, I., et al. (2019). Survey on Frontiers of Language and Robotics. *Adv. Robotics.* 33, 700–730. doi:10.1080/01691864.2019.1632223
- Taniguchi, T., Nakanishi, H., and Iwahashi, N. (2010). Simultaneous Estimation of Role and Response Strategy in Human-Robot Role-Reversal Imitation Learning The 11th IFAC/IFIP/IFORS/IEA Symposium. *IFAC Proc. Volumes.* 43, 460–464. doi:10.3182/20100831-4-fr-2021.00081
- Taniguchi, T., Nagai, T., Nakamura, T., Iwahashi, N., Ogata, T., and Asoh, H. (2016). Symbol Emergence in Robotics: A Survey. *Adv. Robotics.* 30 (Issue 11), 706–728. doi:10.1080/01691864.2016.1164622
- Taniguchi, T., Nakamura, T., Suzuki, M., Kuniyasu, R., Hayashi, K., Taniguchi, A., et al. (2020). Neuro-Serket: Development of Integrative Cognitive System Through the Composition of Deep Probabilistic Generative Models. *New Generation Comput.* 38, 1–26. doi:10.1007/s00354-019-00084-w
- Taniguchi, T., Ugur, E., Hoffmann, M., Jamone, L., Nagai, T., Rosman, B., et al. (2018). Symbol Emergence in Cognitive Developmental Systems: a Survey. *IEEE Trans. Cogn. Developmental Syst.* 11, 494–516. doi:10.1109/TCDS.2018.2867772

- Vedaldi, A., and Fulkerson, B. (2010). "VLFeat: An Open and Portable Library of Computer Vision Algorithms," in ACM International Conference on Multimedia, New York, NY, 1469–1472.
- Wächter, M., and Asfour, T. (2015). Hierarchical Segmentation of Manipulation Actions Based on Object Relations and Motion Characteristics. *Int. Conf. Adv. Robotics.*, 549–556. doi:10.1109/icar.2015.7251510
- Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., et al. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* 13, 600–612. doi:10.1109/tip.2003.819861
- Wu, M., and Goodman, N. (2018). "Multimodal Generative Models for Scalable Weakly-Supervised Learning," in *Advances in Neural Information Processing Systems 31*. Editors S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Montreal, Canada: Curran Associates, Inc.), 5575–5585.
- Xie, J., Girshick, R., and Farhadi, A. (2016). "Unsupervised Deep Embedding for Clustering Analysis," in International conference on machine learning, New York City, NY, 478–487.
- Yang, B., Fu, X., Sidiropoulos, N. D., and Hong, M. (2017). "Towards K-Means-Friendly Spaces: Simultaneous Deep Learning and Clustering," in International conference on machine learning, Sydney, Australia (PMLR), 3861–3870.
- Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning Deep Features for Scene Recognition Using Places Database. *Adv. Neural Inf. Process. Syst.* 27, 487–495.
- 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.
- Copyright © 2021 Kuniyasu, Nakamura, Taniguchi and Nagai. 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.

APPENDIX 1: NETWORK STRUCTURE OF JMVAE

In this section, we describe the network structure of the JMVAE used in the experiment as a comparison method. **Figure 12** presents the network structure of the JMVAE. The encoder and decoder for the images have the same structure as the MNVAE; that is, the encoder has three convolutional layers

and FC layers, and the decoder has an FC layer and seven deconvolutional layers. The encoder and decoder for the words are composed of an FC layer (512 nodes). The structure of the joint encoder is shown in **Figure 12A** and multimodal values can be obtained through an FC layer (512 nodes), the input of which is the concatenated values of extracted features of images and words. The number of dimensions of the latent variables was set to 32, as in the MNVAE.





An Analysis of Personalized Learning Opportunities in 3D VR

Ildikó Horváth *

VR Learning Center, Széchenyi István University, Győr, Hungary

OPEN ACCESS

Edited by:

Carl Vogel,
Trinity College Dublin, Ireland

Reviewed by:

Borbála Berki,
Széchenyi István University, Hungary
Attila Kovari,
University of Dunaújváros, Hungary
György Molnár,
Budapest University of Technology
and Economics, Hungary
Erzsebet Toth,
University of Debrecen, Hungary

***Correspondence:**

Ildikó Horváth
horvath.ildiko@ga.sze.hu

Specialty section:

This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

Received: 28 February 2021

Accepted: 24 August 2021

Published: 20 September 2021

Citation:

Horváth I (2021) An Analysis of
Personalized Learning Opportunities in
3D VR.
Front. Comput. Sci. 3:673826.
doi: 10.3389/fcomp.2021.673826

Due to its constantly developing technological background, VR and AR technology has been gaining increasing popularity not just in industry or business but in education as well. Research in the field of Cognitive Infocommunications (CogInfoCom) shows that using existing digital technologies, online collaboration and cooperation technologies in 3D VR supports cognitive processes, including the finding, processing, memorization and recalling of information. 3D VR environments are also capable of providing users with a much higher level of comprehension when it comes to sharing and interpreting digital workflows. The paper presents a study carried out with the participation of 90 students. The aim of this study is to investigate how the application of 3D VR platforms as personalized educational environments can also increase VR learning efficiency. Besides considering participants' test performance, metrics such as results on visual, auditory and reading-based learning tests for information acquisition, as well as responses on Kolb's learning styles questionnaires are taken into consideration. The participants' learning styles, information acquisition habits were also observed, allowing us to create and offer a variety of learning pathways based on a variety of content types in the 3D VR environment. The students within the study were divided into two groups: a test group receiving personalized training in the MaxWhere 3D VR classroom, and a control group that studied in a general MaxWhere 3D VR space. This research applies both quantitative and qualitative methods to report findings. The goal was to create adaptive learning environments capable of deriving models of learners and providing personalized learning experiences. We studied the correlation between effectiveness of the tasks and Kolb's learning styles. The study shows the major importance of choosing the optimal task type regarding each Kolb learning style and personalized learning environment. The MaxWhere 3D spaces show a high potential for personalizing VR education. The non-intrusive guiding capabilities of VR environments and of the educational content integrated in the 3D VR spaces were very successful, because the students were able to score 20 percent higher on the tests after studying in VR than after using traditional educational tools. Students also performed the same tasks with 8-10 percent faster response times.

Keywords: cognitive infocommunications, learning efficiency, affective learning, personalization, multi-sensory education, VR learning

1 INTRODUCTION

Research in the field of Cognitive Infocommunications (CogInfoCom) show that using existing digital technologies in 3D VR supports cognitive processes, including the finding, processing and memorization of information (Baranyi et al., 2015; Baranyi and Csapó, 2012).

In the past few years the VR Learning Center Research Lab at Széchenyi István University has been investigating the effects of 3D VR spaces on cognitive capabilities, based on the methodologies of CogInfoCom, and is actively working together with other institutions of higher education to test new ideas within VR-based training and to disseminate relevant results.

Our main studies have shown that:

In the area of effectiveness:

- Digital workflows can be reduced by at least 50 percent in 3D VR spaces Lampert et al. (2018).
- Workflows can be reduced with 30 percent fewer user operations and 80 percent fewer machine operations in 3D VR Horváth and Sudár (2018).
- The rate of information recall is 50 percent higher Berki (2018a).
- The MaxWhere 3D platform increases the effectiveness of online collaborative project management software—requiring 72 percent less elementary user operations, and 80 percent less time spent on overview-related tasks Horváth (2019d).

In the area of visual attention and memory:

- The fixation parameters of eye and hand motion are in negative correlation with visual attention, while the distance between the focus of visual attention and the mouse cursor's motion are not correlated to each other Kovari et al. (2020).
- The effectiveness of VR advertisements was higher than classic web-based ads as demonstrated by a questionnaire based free recall test in a 3D space Berki (2018a).
- There is no significant relationship between the effectiveness of completing a task in MaxWhere VR and the spatial memory (measured by the Corsi-task) and the mental rotation ability Berki (2019).

There have been a number of positive results regarding the capabilities of 3D VR for enhancing teaching activities (Csapó et al., 2018; Kvasznicza, 2017; Kvasznicza et al., 2019; Kövecses-Gösi, 2019; Katona and Kovari, 2018; Horvath, 2016; Kuczmann and Budai, 2019; Markopoulos et al., 2020; Botha-Ravyse et al., 2020; Biró et al., 2017; Orosz et al., 2019) the effectiveness of new teaching methodologies (Horváth, 2019b; Kövecses-Gösi, 2019; Kővári, 2019; Horváth, 2019c; Sztahó et al., 2019; Vicsi et al., 2000, Mathability Baranyi and Gilányi, 2013; Chmielewska, 2019) or eye-tracking studies (Hercegfí et al., 2019; Hercegfí and Köles, 2019; Hercegfí et al., 2019) as well.

Most traditional teaching techniques share information through visual or auditory channels. At the same time, multisensory educational techniques and strategies are utilized to engage students at multiple levels (Shams and Seitz, 2008). Using a multisensory teaching technique means that students

are capable of learning in multiple ways. The great advantage of multisensory teaching methods is that they activate more regions of both hemisphere so linking to more senses makes knowledge more memorable and durable.

VR education basically applies a multisensory teaching strategy and exploits the great advantages of 3D. It provides visual, audiovisual and multisensory websites for students besides text based 1st order workflows (Lampert et al., 2018; Horváth and Sudár, 2018) to a multidirectional VR learning content. The goal of multisensory stimulation is to decrease cognitive load.

1.1 Possibilities of Multisensory Teaching in Maxwhere 3D VR Spaces

Our experience in the world incorporates multisensory stimulation. Visual and auditory information are integrated in performing many tasks that involve localizing and tracking moving objects (Shams and Seitz, 2008). Csapó and colleagues 2012 article unfolds how to design multi-sensory signals, which are capable of communicating high-level information to users (Csapó et al., 2012).

VR education can utilize the simultaneous use of visual, auditory, and indirectly (use of mouse, keyboard, touchpad, touch screen and CogiNav technology) kinesthetic (motion-related) channels in multisensory teaching.

3D VR technology provides interactivity by responding to movements. Interactivity is the ability to control events in the simulation by using ones body or several input peripheral devices movements which in turn initiates responses in the simulation as a result of these movements (Christou, 2010).

The content of the smart boards can be selected as desired in the MaxWhere 3D VR spaces. Content can be static or dynamic, and can include text, 3D models, online graphic organizers, calendars, project management software, and more. By highlighting and marking up words, modifying the colors of the smartboard frames, or by changing colors from static to blinking can provide visual support for focus of attention and memorization (Figure 1). In this case, an involuntary action activates the user's attention and the visual memorization. This activity is built up piece by piece in order to facilitate the synthesis of the student's learning. The videos related to the educational topic, which have the advantage of being multisensory on the one hand because they contain both visual and verbal elements and on the other hand presenting information holistically rather than analytically, as in most textbooks and teaching materials.

The 3D objects give students a chance to become more visually immersed and engaged in educational content. 3D objects can be walked around, you can also look inside the object with a movement that is not possible in reality. Disassembled, stackable 3D objects also provide a kinesthetic effect. The reception and retention of information gained by vision may be affected by difficulties in tracking or visual processing. It is important to distinguish between textual and pictorial information. For most users, processing the information they read represents a greater cognitive strain. On the other hand, there are those for whom texts are a more explicit source of information.



FIGURE 1 | Visual supported smartboards in MaxWhere 3D VR.

In MaxWhere spaces, auditive information can also be a source of information. Its exclusive use is not typical, as VR spaces add more to the user's work or learning process with a 3D visual experience. As part of multisensory training, VR also provides auditory information. Sometimes, the subject's auditory processing may be more stressful. Considering students' information acquisition preferences and using multiple senses together can provide a solution to individual problems.

The feeling of natural movement in VR space is provided by MaxWhere 3D VR cognitive navigation (CogiNav) technology. The CogiNav technology operates by taking into consideration the context of the camera (or more generally, the avatar) and the object that it is looking at, using those two pieces of information to deduce the intentions of the user, and finally mapping those intentions onto the limited degrees of freedom of the input device and the characteristics of the camera movements. Thus, although the input dimensions of the input device are limited, their function changes through context in a way that fits naturally with users' expectations from the physical world. The seamless controls for spatial navigation allows it to be used as a communication tool built primarily on 3D interactions and 3D metaphors. Spatial awareness helps to connect visual, auditory, and movement memories to a place, piece of equipment, environment, and involves the delivery of an accurate workflow.

The user's having to expend less effort means that the user feels less disturbed in his or her primary activities while still being able to perceive new information that he or she may find useful. More specifically, in a 3D space with smartboards all along the walls of the environment, users will recognize key changes in the contents of those smartboards even when they are not actively looking at them. A previous study showed, that users could benefit more from the supplementary information in the desktop VR than in a 2D browser if it appeared in their visual field Berki (2018b).

1.2 Concept of Personalization in VR Education

In addition to cognitive factors (Vogel and Esposito, 2020), it is important to consider that people are complex individuals

with different personality types, learning preferences, aptitudes and attitudes toward learning. Each student has different and consistent preferred ways of perception, organization and retention. Some people tend to pick up information better when it is presented verbally, others when it is presented visually and still others prefer reading. These learning styles are good indicators of how students respond to their learning environment.

There are many theories - some of them are really critically for learning styles (Pashler et al., 2008; Coffield et al., 2004) - a move to an active experiential based education explicitly acknowledging different learning styles - has been called transformational learning de Jesus et al. (2006) has been promoted as a more effective alternative to traditional pedagogy (Jacoby, 1996).

Coffield et al. (2004) identified more than 70 learning style models (Coffield et al., 2004). The most influential learning style model is Kolb's model (Kaye et al., 2005; Kaye, 2005). An appeal of Kolb's Experiential Learning Model is its focus on the experiential learning process rather than on fixed learning traits (Turesky and Gallagher, 2011).

However, while there are many theories, little empirical evidence exists for linking such learning styles to 3D VR education.

Students' learning styles never show a unified picture, so if we teach with only one channel, it favors some students while not others. A learning style is also built on the three main channels, which describes the form in which the student can most effectively process the curriculum. Although we are able to receive information through all three channels, the use of a dominant channel and a dominant learning style provides the best results.

In 3D VR learning the kind of analytical thinking and step-by-step information processing representative of earlier times in 2D e-Learning is being replaced by a more comprehensive spatial-visual kind of information processing. It is necessary to select the most effective technology and methodology to support education. However, it is important to realize that learning is not just about memorizing facts, it is also about

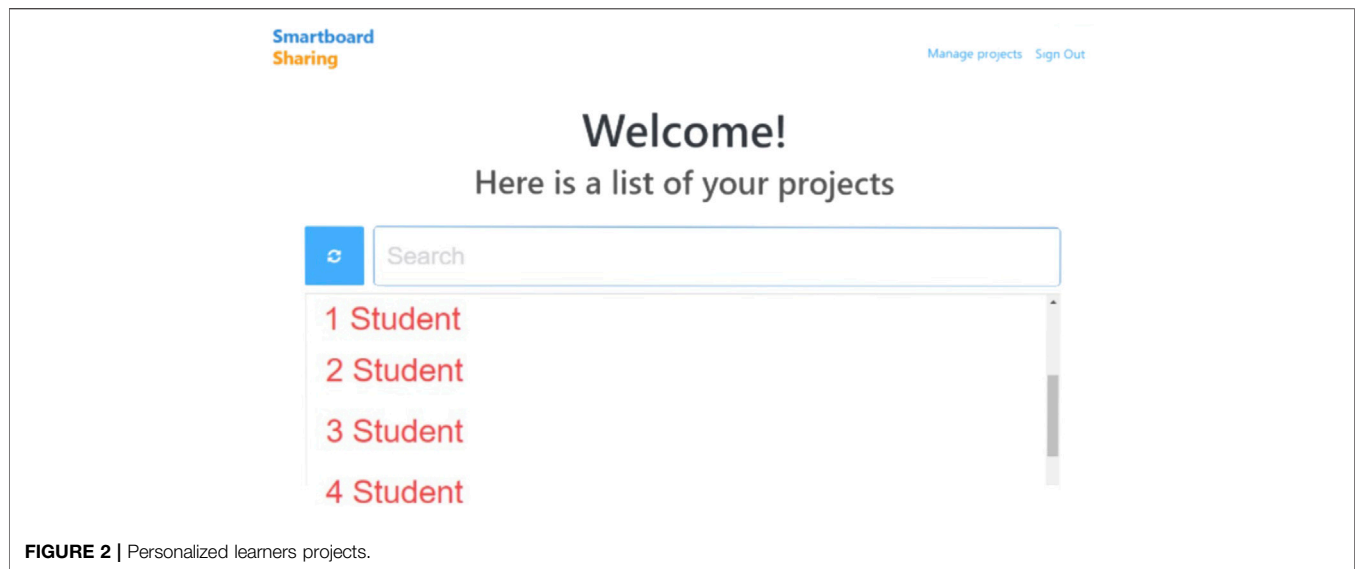


FIGURE 2 | Personalized learners projects.

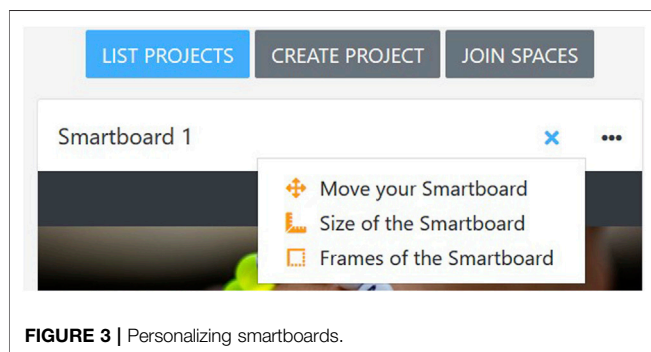


FIGURE 3 | Personalizing smartboards.

learning styles and skills. VR learning activates all learning skills at once.

If we want to personalize VR education, we must pay attention to personalizing the VR learning environment and learning content as well, so learners can choose to engage with the learning materials in the manner that interests them the most. MaxWhere 3D VR platform (www.maxwhere.com) has several opportunity to personalizing the education:

1.2.1 Easily Customizable VR Learning Environment

- Providing personal check-in. All students can make the course “personal” in MaxWhere’s learning spaces. They can capture their name as part of the registration process. Moreover, they can find their own name on the control panel without personal login if the lecturers create personalized educational projects for students (Figure 2). Personalize the learner helps the learning motivation, mainly when we use the name throughout the course.
- The developed Smartboard Sharing technology connecting to 3D VR spaces, which enables users to create personalized

learning spaces by scaling, moving, turning on and off the boards. (Figure 3).

Personalizing learning sequences. Creating “nonlinear” content - Digital Workflows of the 4th order Lampert et al. (2018); Horváth and Sudár (2018) - in 3D VR allows learners to pick and choose how they will learn. Students can optimize their learning path and make personalized story in 3D VR. To implement personalization to suit students’ learning styles and preferences, the research presented in the next chapter was conducted.

This study looks to address this issue, considering not only test performance - visual, auditory and reading-based learning test and Kolb’s learning styles questionnaire - but also other outcomes of using VR for personalized learning. This paper emphasizes that cooperation of humans in virtual spaces enhances affective mental behaviors and motivations in addition to cognitive abilities. The survey was based on an empirical survey and direct measurements with the participation of 90 students (45 students in an experimental group and 45 students in a control group), in September 2019. The study was carried out with the participation of BSc electrical engineering students at the Faculty of Engineering and Information Technology of the University of Pécs, and BSc students of educational sciences at the Faculty of Apáczai Csere János at the Széchenyi István University. Participation in this research was voluntary and conducted with informed consent of the participants.

The paper is structured as follows. The first part draws attention to MaxWhere’s multisensory educational capabilities and the concept of personalization. In the second section, materials and methods are presented. In section 3, the research results of this paper are summarized. Finally, a discussion is provided on how a framework for personalized affective learning in VR education could be implemented.

2 MATERIALS AND METHODS

2.1 Research Background

During the COVID era, the need for personalized learning in online education has intensified. The goal is to ensure independent and successful learning. At the same time, independence does not mean the absence of consultation or space to face meetings, but it means the creation of the possibility of independent or interactive knowledge acquisition without continuous supervision. VR education could be designed several ways on the MaxWhere 3D VR platform, which includes multiple learning preferences so students can choose to relate to the curriculum in a way that interests them. Personalized learning is goals to facilitate the academic success of each student by first determining the learning needs, preferences of individual students, and then providing learning space, contents and methods that are customized for each student. To accomplish this goal we observe the students learning styles, information acquisition habits, create and offer students a variety of learning pathways and contents. Since the 1970s, educational researchers have attempted to categorize the different ways in which people learn and retain information and concepts. “Cognitive styles” or “learning styles” have been defined as “self-consistent, enduring individual differences in cognitive organization and functioning” (Ausubel et al., 1968).

There are multiple factors that impact students’ ability to learn, including age, digital and cultural background beside the learning styles and habits. The results for these factors have been published previously (Horváth, 2019a; Komlósi and Waldbuesser, 2015). In this study a test identifying preferred information acquisition habits and the Kolb’s Learning Style test (Kolb, 1981) were applied.

2.1.1 Test for Mapping Preferred Information Acquisition Habits

The learning style is made up of a combination of several factors, such as:

- What is the easiest way to receive information?
- How we organize and process information?
- What conditions do we need to help pick up and store information?
- How can we retrieve information?

The goal of this test was to created in order to assess the students for preferred information acquisition habits. The relevance of this test in this experiment was that the students were grouped by visual, auditive and reading categories based on this test results.

To test the visual skills, students saw 10 pictures in a row. They did not take notes. This was followed by a task. This task diverted their thoughts from the images they saw. The task was to answer the question: “If Anna’s birthday was on the 60th day of the year, and the year was not a leap year, then on which day was she born?”

After answering the task, the words the students remembered had to be written down based on the pictures they saw. They had 1 min to complete the task.

The auditive test was similar. In this case 10 words were heard, followed by a logical task, after which the words had to be recalled from memory.

In the case of the reading test, the students saw 10 words. Next task was: “Yesterday was the 3rd day after Monday. What day will be the day after tomorrow?”

After answering, participants had to write down the words they memorized.

Mapping the kinesthetic learning style was not part of the study. This is because the user can move in 3D using a mouse or touchpad. Its seamless controls for spatial navigation allows it to be used as a communication tool built primarily on 3D interactions and 3D metaphors.

2.1.2 Kolb’s Learning Style Test for Information Processing Preferences

David A Kolb an American social psychologist and learning researcher, developed a test to determine learning style. Kolb separated four different learning style. According to his theory learning is a cyclically repetitive process, where experience, observation, thinking and application stages are separated. The learning cycle depends on a double basis: the recording, obtaining and processing the information. Kolb identified four different learning styles (Kolb, 1981):

- Accommodating: The information that is retained is based on experience and is processed by an active experimenter. This is an action oriented, flexible, risk taking and quickly applied way of learning.
- Diverging: Specific information is gathered through experiences of a reflective observer. Information in this case is processed through observation and understanding.
- Assimilating: Through observational experience, this learning style processes abstract information. The assimilating style is logical, theory oriented and creates coordinated models.
- Converging: This style uses abstract information in a practical, action-oriented, purposeful, and planning-oriented way.

Kolb created a 12 question test to determine the learning style, which can be an important step in our work of self-knowledge. We can become aware of the characteristic way of obtaining and processing information during learning and working. This awareness helps us balance the educational goal and learning materials and methods or understand and accept the learning style of students. The positive outcome of acceptance, understanding and apprehension of conflicts is that we are more aware of our own learning style.

We were looking for opportunities to implement personalized education in 3D VR spaces.

Before completing the test, students were asked what type of learning style they classified themselves. The aim of the question was to find out whether they had prior knowledge of their own learning processes and cognitive abilities.

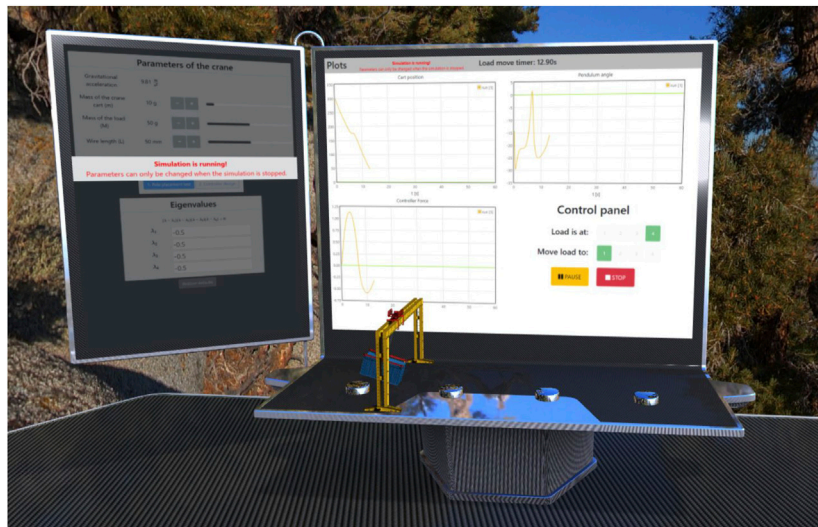


FIGURE 4 | Interactive VR learning space with 3D object.

To fulfill the test all students received the curriculum in MaxWhere 3D VR space and all students were given equal time to learn. The students in the focus group received a personalized curriculum, while the control group got multisensory curriculum.

2.2 The Personalized Learning Curriculum Description

In terms of gathering information for visual type learners, the curriculum was created with graphs, flowcharts, images, videos, and 3D objects. Keywords are highlighted in the text information. Tasks always included, “Make a picture/graph etc.!” instruction as well.

For the auditory type of students, we provided the lectures with audio recording, online meeting applications, they had the opportunity to discuss the topic, request additional oral information and audio videos were added to the theoretical material. Tasks included, “Tell/Summarize!” instructions. The recordings could also be uploaded using Smartboard Sharing, so there was also an opportunity for teacher feedback.

The educational content for students who preferred to read written text included presentations, additional notes, articles, and written web content. Their job was to create and upload their own notes and summary on the Smartboard Sharing.

In the case of mixed-type students, the curriculum was created by blending the above. VR activities could be designed to include multiple learning methods, so according to the results of the Kolb learning style test, students of the Accommodating type were given more interactive contents an activity task with quick feedback. We gave a large number of articles and websites to students belonging to the Diverging type. Their task is with instructions: “collect, organize, formulate somethings” we have given. Students who belong to the Assimilating type received more information based on the principle or rule, than what the curriculum focuses.

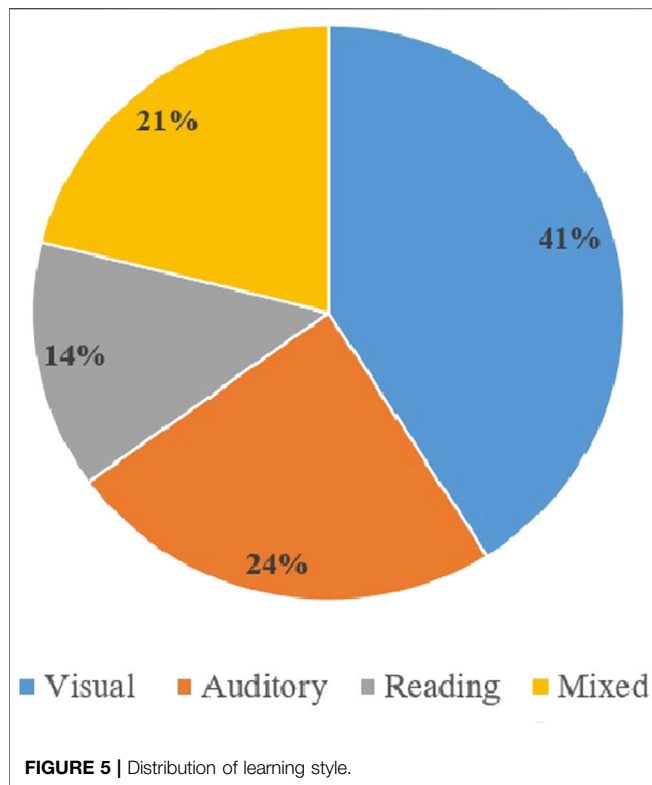
In addition to receiving presentations, articles, or visions, their tasks were to create models, other types of uses. The students belonging to Converging learning style are action oriented and purposeful, so the optimal educational methods are the project based educational or Edu-Coaching methods. They were asked to provide a learning goal before beginning the learning process. Why is it important for them to acquire that knowledge? For Converging learning style students, we can often use Laboratory spaces that also included simulations that worked with 3D objects (Figure 4).

2.3 Participant Recruitment

This research was reviewed and approved by both the ethical review committee for Doctoral School of Multidisciplinary Engineering Sciences (MMTDI) at the Széchenyi István University and the ethical review committee for University of Pécs, Faculty of Engineering and Information Technology, Department of Electrical Network.

At the first lesson of the course the participants were informed in detail about the purpose, conditions and background of the research. This detailed information was: “The aim of the study is to observe the categorization of learning preferences and styles. During the task, participants see pictures, hear and read words, and then describe what they remember. After the examination, a questionnaire (Kolb’s learning styles questionnaire) is also completed. The test above has no adverse consequences. The test takes approximately 30 min. The study can be stopped at any time without consequences, in which case the data already recorded will not be evaluated. After the study, participants learn in a personalized educational environment and curricula. Researchers measure their learning outcomes.”

Participants were volunteers. Consenting participants made assertions that they have no known neurological or psychiatric disorders. They declared that their vision and hearing is intact or corrected to normal. I have no dyslexia, dysgraphia or other



learning difficulties. The manuscript has the English text of the consent form as an **Supplementary Appendix**. Participants had declared consent below:

“I declare that I have received satisfactory information about the goals and nature of the study.

I declare that I am participating in this study voluntarily.

Neither I nor my relatives requested or received financial compensation for my contribution to the investigations.

I acknowledge that the personally identifiable information will be treated confidentially by the study manager and will not be disclosed to anyone other than those involved in the conduct of the experiment.

I agree that non-identifiable data included in the study should be made available to other researchers.

I acknowledge that the test data are for research and non-diagnostic purposes only.

I do not request such an opinion after the tests have been performed.

I, the undersigned (name), agree that I will participate in the examinations of Széchenyi István University on (day) September 2019.”

2.4 Statistical Analysis

After developing personalized VR educational spaces and learning content, the 45 students (in experimental group) included in the study continued their personalized learning. The control group also included 45 students from the same age group. They also received the educational contents in 3D VR space, but their learning style was not determined and applied.

90 (56 male and 34 female) students participated in the study. The average age of participants was 20.188 (between 19 and 24 years).

Levene’s test was applied to assess the equality of variances for a variable calculated for experimental and control group, before we perform the independent *t*-test. The null hypothesis formed in the Levene’s test is that the groups we are comparing have equal variance.

The Independent Samples *t* Test compares the means of two independent groups - the experimental and the control group - in order to determine whether there is statistical evidence that the means of total score and response time are significantly different in our two groups.

3 RESEARCH RESULTS

3.1 Learning Style Test

In terms of visual, auditory and reading-based learning test, 42 percent of students preferred visual information, 24 percent of them the auditory, 13 percent of them the reading based and 21 percent the mixed (visual and reading, visual and auditory, visual-auditory and reading) information (**Figure 5**).

A surprisingly high proportion of the students were unable to answer the question (47 percent), or gave a wrong answer the question (15.5 percent): “What kind of learning style do you prefer?” The answer was wrong if the student did not name the own real learning test. Students stated that so far they have not had the opportunity to choose between types of educational materials, so they have not addressed this issue.

3.2 Kolb’s Learning Style Test

According to Kolb’s learning test results, the distribution of students’ learning styles is shown in **Figure 6**.

Logical thinking, modelling and design are connected with the assimilating learning type, which is mainly present in the field of engineering.

3.3 Statistical Analysis

The learning performance of the two groups were examined. To test the hypothesis that the experimental group ($N = 45$), who got the personalized curriculum for learning and the control group ($N = 45$) were associated with statistically significant different mean response time and total score, an independent samples *t*-test was performed. Additionally, the assumption of homogeneity of variances was tested and satisfied via Levene’s Test $F = 0.375$, $p = 0.542$ to total Score and $F = 0.387$, $p = 0.532$ to response time.

Based on statistical analysis, we accept the null hypothesis of Levene’s test and conclude that the variance in total score and in response time variance are accepted as equal. Levene’s statistic response time is shown in **Figure 7** and the total score is shown in **Figure 8**.

To compare the results of the experimental and control group an independent sample *t*-test was used. The null hypothesis of the independent samples *t*-test is that the means of total score and

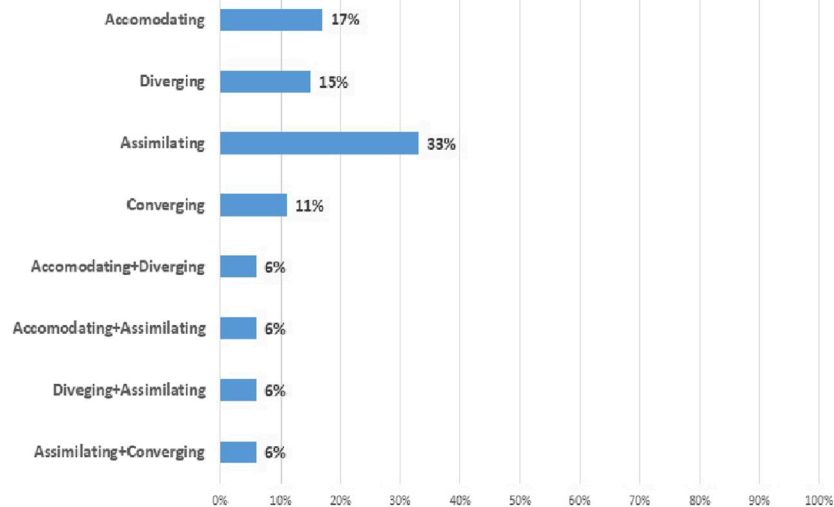


FIGURE 6 | Distribution of learning style based on Kolb's test.

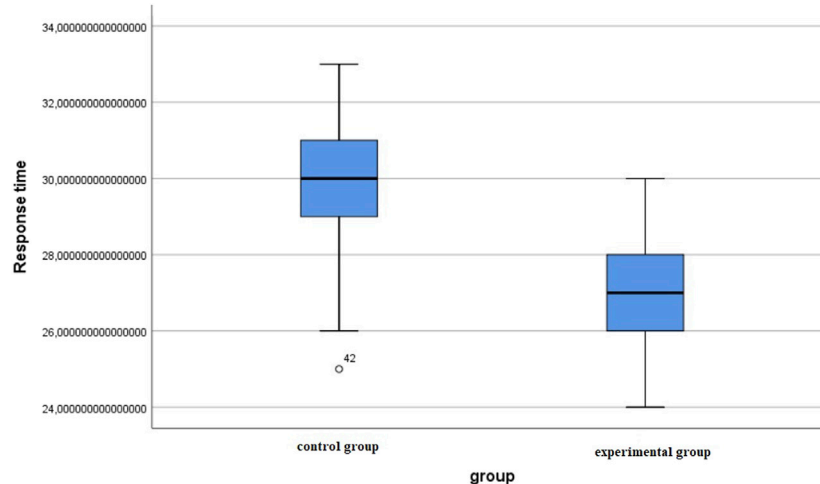


FIGURE 7 | Response time.

response time are equal in the two groups. The mean total score in the control group was 38.73 (SD = 8.29) and 48.44 in the experimental group (SD = 7.3). The difference between these means are statistically significant based on the results of the *t*-test: $t(88) = -5.898, p < 0.01$. The means response time in the control group was 29.58 min (SD = 1.69) and 27.02 min (SD = 1.29) in the experimental group. This difference is statistically significant based on the results of the *t*-test: $t(88) = 8.082, p < 0.01$. This means that the test score in the experimental group, who received personalized educational content, was significantly higher, also their response times were significantly faster.

The correlation between the test score and the characteristic score was examined. The characteristic score of Kolb's test represents the strength of group membership. There was a positive correlation between personalized educational content

and tests' score ($r = 0.784, p < 0.01$). Correlation is significant on the level of 0.01. The test score in the focus group (personalized education) was on average 20 percent higher than in the control group, with 8–10 percent faster response times. There was a significant performance improvement for memory recall type tasks. Students in the focus group remembered more than 30 percent more factual data. In case the question contains the initial information (facts), and the task was to process them. The difference here was not significant. We studied the correlation between effectiveness of the tasks and Kolb's learning styles. The study shows the major importance of choosing the optimal task type regarding each Kolb learning style and personalized learning environment. According to the results of the correctional investigations, the Accommodating learning style shows a negative significant correlation ($r = -.783; p < 0.01$)

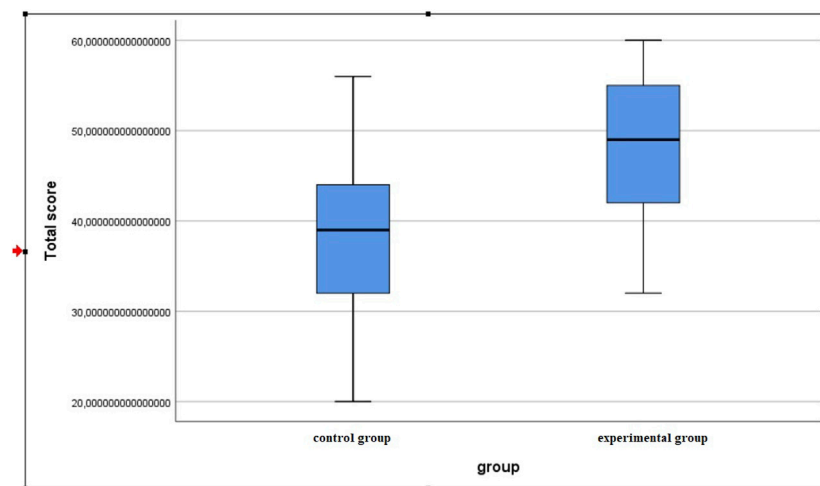


FIGURE 8 | Total score.

with learning results of long descriptive or long and tedious explanatory tasks. This result concerns students who are a part of the control group, so they studied in a non-personalized learning environment.

The other definitive result was found for the students with a Diverging learning style. In this case, the active presentations were less successful and caused difficulties for these students ($r = -0.688$; $p < 0.01$).

4 DISCUSSION

Traditional educational methods and their tools and sources, such as presentations or textbooks typically guide the students to follow fixed sequences to related to their learning process. In order to integrate the information in the knowledge system, the students have to go through the process of information processing which cannot eliminate one or the other phase. However, the inappropriate courseware leads to learner cognitive overload or disorientation during learning processes, thus reducing learning performance (Chen, 2008). We see, that the Web-based instruction researchers have given considerable attention to flexible curriculum sequencing control to provide adaptable, personalized learning programs (Brusilovsky and Vassileva, 2003; Hübscher, 2000; Papanikolaou et al., 2002; Lee, 2001). In an educational adaptive hypermedia system, an optimal learning path aims to maximize a combination of the learner's understanding of courseware and the efficiency of learning the courseware (Hübscher (2000)). Nowadays, most adaptive/personalized tutoring systems (Lee, 2001; Papanikolaou et al., 2002; Tang and McCalla, 2003) consider student preferences, when investigating learner behaviors for personalized services.

However, while there are many theories little empirical evidence exists for linking such learning styles to 3D VR education. Our previous papers focus on information finding, filtering, reception, processing and applications.

This paper investigates the opportunity of personalization in 3D VR. The main point here is that in order for learning to be effective. Learning is by necessity a multi-faceted process. Students must be able to recall facts, definitions, the relationships and to create new relevant information. 3D environment is absolutely effective because the human brain was formed in an evolutionary environment where the recognition and recollection of spatial relationships and locations was of paramount importance (much more so than the rote memorization of non-spatial concepts) (Logie, 1986; Rodriguez et al., 2002; Csapó et al., 2018).

If we want to personalize VR education the best way is the categorization and optimizing of VR educational environments and contents. The results of the two types of learning style tests helped us to design this. At first the results of the two types of learning tests were analyzed and individual educational content was developed for each participant in experimental group. Learning style test for preferred information acquisition habits provided us with relevant information for creating individually preferred content. Kolb learning test results were used in connection with information processing. For example, regarding the Accommodating learning style in the case of students in the personalized group, the most successful tasks were the ones which were short, substantial and concise. The students with a Diverging learning type were most efficient in tasks where they could do analytical work and work alone. There isn't such a strong correlation between the other two learning styles and the several tasks.

To implement an automated and personalized VR educational system, we first need a database in which the curriculum contents are stored in the form of visual, audio visual and text based data. After that we have to determine the individual's primary information gathering preference by using sensors or other techniques. For user interaction there are several options: head tracking, eye tracking, motion tracking etc. If the information obtained by eye movement tracking or other sensory based device

is the same as the results of classical information acquisition preference tests, then starting from a mixed information base, after a few minutes the personalized information content can be loaded into the smart boards automatically, from the database behind MaxWhere VR system.

Our original idea was based on the examination based on the tracking of eye movement. However, due to the pandemic, we were unable to perform this measurement, which is part of our work for further research and development.

5 CONCLUSIONS

With a personalized learning experience, every student can enjoy an unique educational approach in 3D VR that is fully tailored to his or her individual abilities and needs. This can directly increase students' motivation and reduce their likelihood of dropping out. It could also help achieve a better understanding of each student's learning process. Therefore, the education system could benefit from the application of tests for different types of learning styles. This paper examined learning effectiveness as it relates to the use of personalized learning content. The goal was to create adaptive learning environments capable of deriving models of learners and providing personalized learning experiences.

The MaxWhere 3D spaces show a high potential for personalizing VR education. The non-intrusive guiding capabilities of VR environments and of the educational content integrated in the spaces were successful, because the students were able to score 20 percent higher on the tests after studying in VR than after using traditional educational tools.

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.

REFERENCES

- Ausubel, D. P., Novak, J. D., and Hanesian, H. (1968). *Educational Psychology: A Cognitive View*. New York: Holt Rinehart & Winston.
- Baranyi, P., and Csapó, Á. (2012). Definition and Synergies of Cognitive Infocommunications. *Acta Polytechnica Hungarica* 9, 67–83.
- Baranyi, P., Csapó, A., and Sallai, G. (2015). *Cognitive Infocommunications (CogInfoCom)*. New York: Springer.
- Baranyi, P., and Gilányi, A. (2013). "Mathability: Emulating and Enhancing Human Mathematical Capabilities," in 2013 IEEE 4th international conference on cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 555–558. doi:10.1109/coginfocom.2013.6719309
- Berki, B. (2018a). "Better Memory Performance for Images in Maxwhere 3d Vr Space Than in Website," in 2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 000281–000284. doi:10.1109/coginfocom.2018.8639956
- Berki, B. (2018b). "Desktop Vr and the Use of Supplementary Visual Information," in 2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 000333–000336. doi:10.1109/coginfocom.2018.8639925

ETHICS STATEMENT

This research was reviewed and approved by both the ethical review committee for Doctoral School of Multidisciplinary Engineering Sciences (MMTDI) at the Szechenyi Istvan University and the ethical review committee for University of Pecs, Faculty of Engineering and Information Technology, Department of Electrical Network. Written informed consent to participate in the study was provided by the participants

AUTHOR CONTRIBUTIONS

IH contributed to conception and design of the study. IH organized the database. IH performed the statistical analysis. IH wrote the manuscript. The author read, and approved the submitted version I have an Agreement between IH ("Me") and IEEE ("IEEE") consists of my license details and the terms and conditions provided by IEEE and Copyright Clearance Center.

ACKNOWLEDGMENTS

The content of this manuscript has been Horváth (2020) presented in part at the 11th IEEE International Conference on Cognitive InfoCommunications. I would like to thank the IEEE for permission to reuse the published conference paper. This work was supported by the FIEK program (Center for cooperation between higher education and the industries at the Széchenyi István University, GINOP-2.3.4-15-2016-00003).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcomp.2021.673826/full#supplementary-material>

- Berki, B. (2019). Does Effective Use of Maxwhere Vr Relate to the Individual Spatial Memory and Mental Rotation Skills? *Acta Polytechnica Hungarica* 16, 41–53. doi:10.12700/aph.16.6.2019.6.4
- Biró, K., Molnár, G., Pap, D., and Szűts, Z. (2017). "The Effects of Virtual and Augmented Learning Environments on the Learning Process in Secondary School," in 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 000371–000376. doi:10.1109/coginfocom.2017.8268273
- Botha-Ravyse, C., Fr, T. H., Luimula, M., Markopoulos, P., Bellalouna, F., et al. (2020). "Towards a Methodology for Online Vr Application Testing," in 2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 000295–000298. doi:10.1109/coginfocom50765.2020.9237893
- Brusilovsky, P., and Vassileva, J. (2003). Course Sequencing Techniques for Large-Scale Web-Based Education. *Int. J. Contin. Eng. Edu. Lifelong Learn.* 13, 75–94. doi:10.1504/ijceell.2003.002154
- Chen, C.-M. (2008). Intelligent Web-Based Learning System with Personalized Learning Path Guidance. *Comput. Edu.* 51, 787–814. doi:10.1016/j.compedu.2007.08.004
- Chmielewska, K. (2019). "From Mathability to Learnability," in 2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 513–516. doi:10.1109/coginfocom47531.2019.9089891

- Christou, C. (2010). "Virtual Reality in Education," in *Affective, interactive and cognitive methods for e-learning design: creating an optimal education experience* (Pennsylvania: IGI Global), 228–243.
- Coffield, F., Ecclestone, K., Hall, E., and Moseley, D. (2004). *Learning Styles and Pedagogy in post-16 Learning: A Systematic and Critical Review*. New Delhi: Learning and Skills Research Council.
- Csapó, Á., Baranyi, P., and Várlaki, P. (2012). "A Taxonomy of Coginfocom Trigger Types in Practical Use Cases," in 2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 21–25. doi:10.1109/coginfocom.2012.6421990
- Csapó, Á. B., Horvath, I., Galambos, P., and Baranyi, P. (2018). "Vr as a Medium of Communication: from Memory Palaces to Comprehensive Memory Management," in 2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 000389–000394. doi:10.1109/coginfocom.2018.8639896
- de Jesus, H. T. P., Almeida, P. A., Teixeira-Dias, J. J., and Watts, M. (2006). Students' Questions: Building a Bridge between Kolb's Learning Styles and Approaches to Learning. *Edu. Train.* 48, 97. doi:10.1108/00400910610651746
- Hercegi, K., and Köles, M. (2019). Temporal Resolution Capabilities of the Mid-frequency Heart Rate Variability-Based Human-Computer Interaction Evaluation Method. *Acta Polytechnica Hungarica* 16, 95–114. doi:10.12700/aph.16.6.2019.6.7
- Hercegi, K., Komlósi, A., Köles, M., and Tóvölgyi, S. (2019). Eye-tracking-based Wizard-Of-Oz Usability Evaluation of an Emotional Display Agent Integrated to a Virtual Environment. *Acta Polytechnica Hungarica* 16, 145–162. doi:10.12700/aph.16.2.2019.2.9
- Horváth, I. (2019a). "Behaviors and Capabilities of Generation Ce Students in 3d Vr," in 2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 491–494. doi:10.1109/coginfocom.2019.9089988
- Horváth, I. (2019c). "How to Develop Excellent Educational Content for 3d Vr," in 2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 483–490. doi:10.1109/coginfocom.2019.9089916
- Horvath, I. (2016). "Innovative Engineering Education in the Cooperative Vr Environment," in 2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 000359–000364. doi:10.1109/coginfocom.2016.7804576
- Horváth, I. (2019d). Maxwhere 3d Capabilities Contributing to the Enhanced Efficiency of the Trello 2d Management Software. *Acta Polytechnica Hungarica* 16, 55–71. doi:10.12700/aph.16.6.2019.6.5
- Horváth, I. (2020). "Personalized Learning Opportunity in 3d Vr," in 2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 000425–000430. doi:10.1109/coginfocom.2020.9237895
- Horváth, I., and Sudár, A. (2018). Factors Contributing to the Enhanced Performance of the Maxwhere 3d Vr Platform in the Distribution of Digital Information. *Acta Polytechnica Hungarica* 15, 149–173. doi:10.12700/aph.15.3.2018.3.9
- Horváth, I. (2019b). "The Edu-Coaching Method in the Service of Efficient Teaching of Disruptive Technologies," in *Cognitive Infocommunications, Theory and Applications* (New York: Springer), 349–363. doi:10.1007/978-3-319-95996-2_16
- Hübscher, R. (2000). "Logically Optimal Curriculum Sequences for Adaptive Hypermedia Systems," in *International conference on adaptive hypermedia and adaptive web-based systems* (New York: Springer), 121–132. doi:10.1007/3-540-44595-1_12
- Jacoby, B. (1996). "Service-learning in Today's Higher Education," in *Service-Learning in Higher Education: Concepts and Practices*. Editors B. Jacoby and Associates (San Francisco, CA: Jossey-Bass), 3–25.
- Katona, J., and Kovari, A. (2018). Examining the Learning Efficiency by a Brain-Computer Interface System. *Acta Polytechnica Hungarica* 15, 251–280. doi:10.12700/aph.15.3.2018.3.14
- Kayes, A. B., Kayes, D. C., and Kolb, D. A. (2005). Experiential Learning in Teams. *Simulation & Gaming* 36, 330–354. doi:10.1177/1046878105279012
- Kayes, D. C. (2005). Internal Validity and Reliability of Kolb's Learning Style Inventory Version 3 (1999). *J. Bus Psychol.* 20, 249–257. doi:10.1007/s10869-005-8262-4
- Kolb, D. A. (1981). Learning Styles and Disciplinary Differences. *Mod. Am. Coll.* 1, 232–235.
- Komlósi, L. I., and Waldbuesser, P. (2015). "The Cognitive Entity Generation: Emergent Properties in Social Cognition," in 2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 439–442. doi:10.1109/coginfocom.2015.7390633
- Kovari, A., Katona, J., and Costescu, C. (2020). Quantitative Analysis of Relationship between Visual Attention and Eye-Hand Coordination. *Acta Polytech Hung* 17, 77–95. doi:10.12700/aph.17.2.2020.2.5
- Kövecses-Gösi, V. (2019). "The Pedagogical Project of Education for Sustainable Development in 3d Virtual Space," in 2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 000539–000544. doi:10.1109/coginfocom.2019.9090000
- Kuczmann, M., and Budai, T. (2019). Linear State Space Modeling and Control Teaching in Maxwhere Virtual Laboratory. *Acta Polytechnica Hungarica* 16, 27–39. doi:10.12700/aph.16.6.2019.6.3
- Kvasznica, Z., Kovács, J., Maza, G., and Péli, M. (2019). "Vr based duale education at e. on the win-win-win situation for companies, graduates and universities," in 2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 479–482. doi:10.1109/coginfocom.2019.9089961
- Kvasznica, Z. (2017). "Teaching Electrical Machines in a 3d Virtual Space," in 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 000385–000388. doi:10.1109/coginfocom.2017.8268276
- Kovári, A. (2019). Adult Education 4.0 and Industry 4.0 Challenges in Lifelong Learning. *PedActa* 9, 9–16.
- Lampert, B., Pongrácz, A., Sipos, J., Vehrer, A., and Horvath, I. (2018). Maxwhere Vr-Learning Improves Effectiveness over Clasiccal Tools of E-Learning. *Acta Polytechnica Hungarica* 15, 125–147. doi:10.12700/aph.15.3.2018.3.8
- Lee, M.-G. (2001). Profiling Students' Adaptation Styles in Web-Based Learning. *Comput. Edu.* 36, 121–132. doi:10.1016/s0360-1315(00)00046-4
- Logie, R. H. (1986). Visuo-spatial Processing in Working Memory. *The Q. J. Exp. Psychol. Section A* 38, 229–247. doi:10.1080/14640748608401596
- Markopoulos, E., Markopoulos, P., Laiuori, N., Moridis, C., and Luimula, M. (2020). "Finger Tracking and Hand Recognition Technologies in Virtual Reality Maritime Safety Training Applications," in 2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (New Jersey: IEEE), 000251–000258. doi:10.1109/coginfocom.2020.9237915
- Orosz, B., Kovács, C., Karuović, D., Molnár, G., Major, L., Vass, V., et al. (2019). Digital Education in Digital Cooperative Environments. *J. Appl. Tech. Educ. Sci.* 9, 55–69.
- Papanikolaou, K. A., Grigoriadou, M., Magoulas, G. D., and Kornilakis, H. (2002). Towards New Forms of Knowledge Communication: the Adaptive Dimension of a Web-Based Learning Environment. *Comput. Edu.* 39, 333–360. doi:10.1016/s0360-1315(02)00067-2
- Pashler, H., McDaniell, M., Rohrer, D., and Bjork, R. (2008). Learning Styles. *Psychol. Sci. Public Interest* 9, 105–119. doi:10.1111/j.1539-6053.2009.01038.x
- Rodriguez, F., López, J. C., Vargas, J. P., Broglio, C., Gómez, Y., and Salas, C. (2002). Spatial Memory and Hippocampal Pallium through Vertebrate Evolution: Insights from Reptiles and Teleost Fish. *Brain Res. Bull.* 57, 499–503.
- Shams, L., and Seitz, A. R. (2008). Benefits of Multisensory Learning. *Trends Cognitive Sciences* 12, 411–417. doi:10.1016/j.tics.2008.07.006
- Sztahó, D., Szaszák, G., and Beke, A. (2019). Deep Learning Methods in Speaker Recognition: a Review. *arXiv preprint arXiv:1911.06615*.
- Tang, T. Y., and McCalla, G. (2003). "Smart Recommendation for an Evolving E-Learning System," in *Workshop on Technologies for Electronic Documents for Supporting Learning*, International Conference on Artificial Intelligence in Education, Sydney, Australia, 699–710.
- Turesky, E. F., and Gallagher, D. (2011). Know Thyself: Coaching for Leadership Using Kolb's Experiential Learning Theory. *Coaching Psychol.* 7, 5–14.

- Vicsi, K., Roach, P., Öster, A., Kacic, Z., Barczikay, P., Tantos, A., et al. (2000). A Multimedia, Multilingual Teaching and Training System for Children with Speech Disorders. *Int. J. speech Technol.* 3, 289–300. doi:10.1023/a:1026563015923
- Vogel, C., and Esposito, A. (2020). Interaction Analysis and Cognitive Infocommunications. *Infocommunications J.* 12, 2–9. doi:10.36244/icj.2020.1.1

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 reviewer BB declared a shared affiliation with the author IH to the handling editor at time of review.

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.

Copyright © 2021 Horváth. 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.



Unifying Physical Interaction, Linguistic Communication, and Language Acquisition of Cognitive Agents by Minimalist Grammars

Ronald Römer^{1*}, Peter beim Graben², Markus Huber-Liebl¹ and Matthias Wolff¹

¹ Brandenburg University of Technology Cottbus–Senftenberg, Cottbus, Germany, ² Bernstein Center for Computational Neuroscience, Berlin, Germany

OPEN ACCESS

Edited by:

Anna Esposito,
University of Campania "Luigi
Vanvitelli, Italy

Reviewed by:

Farhan Mohamed,
University of Technology Malaysia,
Malaysia
Javad Khodadoust,
Instituto de Tecnología y Educación
Superior de Monterrey (ITESM),
Mexico

*Correspondence:

Ronald Römer
ronald.roemer@b-tu.de

Specialty section:

This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

Received: 30 June 2021

Accepted: 04 January 2022

Published: 27 January 2022

Citation:

Römer R, beim Graben P,
Huber-Liebl M and Wolff M (2022)
Unifying Physical Interaction,
Linguistic Communication, and
Language Acquisition of Cognitive
Agents by Minimalist Grammars.
Front. Comput. Sci. 4:733596.
doi: 10.3389/fcomp.2022.733596

Cognitive agents that act independently and solve problems in their environment on behalf of a user are referred to as autonomous. In order to increase the degree of autonomy, advanced cognitive architectures also contain higher-level psychological modules with which needs and motives of the agent are also taken into account and with which the behavior of the agent can be controlled. Regardless of the level of autonomy, successful behavior is based on interacting with the environment and being able to communicate with other agents or users. The agent can use these skills to learn a truthful knowledge model of the environment and thus predict the consequences of its own actions. For this purpose, the symbolic information received during the interaction and communication must be converted into representational data structures so that they can be stored in the knowledge model, processed logically and retrieved from there. Here, we firstly outline a grammar-based transformation mechanism that unifies the description of physical interaction and linguistic communication and on which the language acquisition is based. Specifically, we use minimalist grammar (MG) for this aim, which is a recent computational implementation of generative linguistics. In order to develop proper cognitive information and communication technologies, we are using utterance meaning transducers (UMT) that are based on semantic parsers and a *mental lexicon*, comprising syntactic and semantic features of the language under consideration. This lexicon must be acquired by a cognitive agent during interaction with its users. To this aim we outline a reinforcement learning algorithm for the acquisition of syntax and semantics of English utterances. English declarative sentences are presented to the agent by a teacher in form of utterance meaning pairs (UMP) where the meanings are encoded as formulas of predicate logic. Since MG codifies universal linguistic competence through inference rules, thereby separating innate linguistic knowledge from the contingently acquired lexicon, our approach unifies generative grammar and reinforcement learning, hence potentially resolving the still pending Chomsky-Skinner controversy.

Keywords: cognitive agents, cognitive architectures, artificial intelligence, reinforcement learning, interaction and communication, minimalist grammars, semantic representations, utterance meaning transducer

1. INTRODUCTION

Traditionally, the technical replication of cognitive systems is based on cognitive architectures with which the most important principles of human cognition are captured. The best known traditional architectures are SOAR and ACT (Funke, 2006). Such architectures serve to integrate psychological findings in a formal model that is as economical as possible and capable of being simulated. It assumes that all cognitive processes can be traced back to a few basic principles. According to Eliasmith (2013), this means that cognitive architectures have suitable *representational data structures*, that they support the *composition-, adaptation- and classification principle* and that they are autonomously capable to gain knowledge by *logical reasoning and learning*. Further criteria are productivity, robustness, scalability and compactness. In psychological research, these architectures are available as computer programs, which are used to empirically test psychological theories. In contrast, the utility of cognitive architectures in artificial intelligence (AI) research lies primarily in the construction of intelligent machines and the ability to explain their behavior.

Explainability of intelligent machines is closely tied to the idea of a *physical symbol system* (PSS) (Newell and Simon, 1976). A PSS takes physical symbols from its sensory equipment, composing them into symbolic structures (expressions) and transforms them to new expressions (Vera and Simon, 1993). For our approach it is crucial that the symbols or the symbol structures, respectively, can be assigned a meaning and that the transformation of these symbol structures leads to logically processable knowledge on which problem solving is based. In order to build meaningful symbols, the agent must be embedded in a *perception-action-cycle* (PAC) and it must be taken into account that the truthfulness or veridicality of the translation of sensory information into symbolic structures is not necessarily guaranteed due to deceptions or dysfunctions (Bischof, 2009). To overcome these difficulties, more sophisticated agents are able to build a dynamic model of their environment that can be simulated. In this case, the agent has to distinguish between redundancy expectations (model-based prior knowledge) and sensory data (observations), so that information from two sources has to be processed. To achieve veridicality, both pieces of information must be combined in a suitable manner (e.g., through a Bayesian model). The structure required for this kind of information processing is depicted by the inner cognitive loop in **Figure 1** and is denoted as interaction¹. This picture illustrates that autonomous behavior can arise if the interaction process is based on truthful information processing that is controlled by higher-level psychological modules. This includes necessities, motives and acquired authorizations that support well-being. In addition to the associated autonomous interaction scenarios, human-machine communication is another area of application for cognitive agents. In this case, the tool or service character of cognitive agents plays a prominent role. This property allows the user to solve certain tasks or problems through the agent.

For this purpose, linguistic descriptions must be articulated, understood and exchanged. This results in the requirement for a speech-enabled agent and an additional outer cognitive loop (see **Figure 1**) which is denoted as communication².

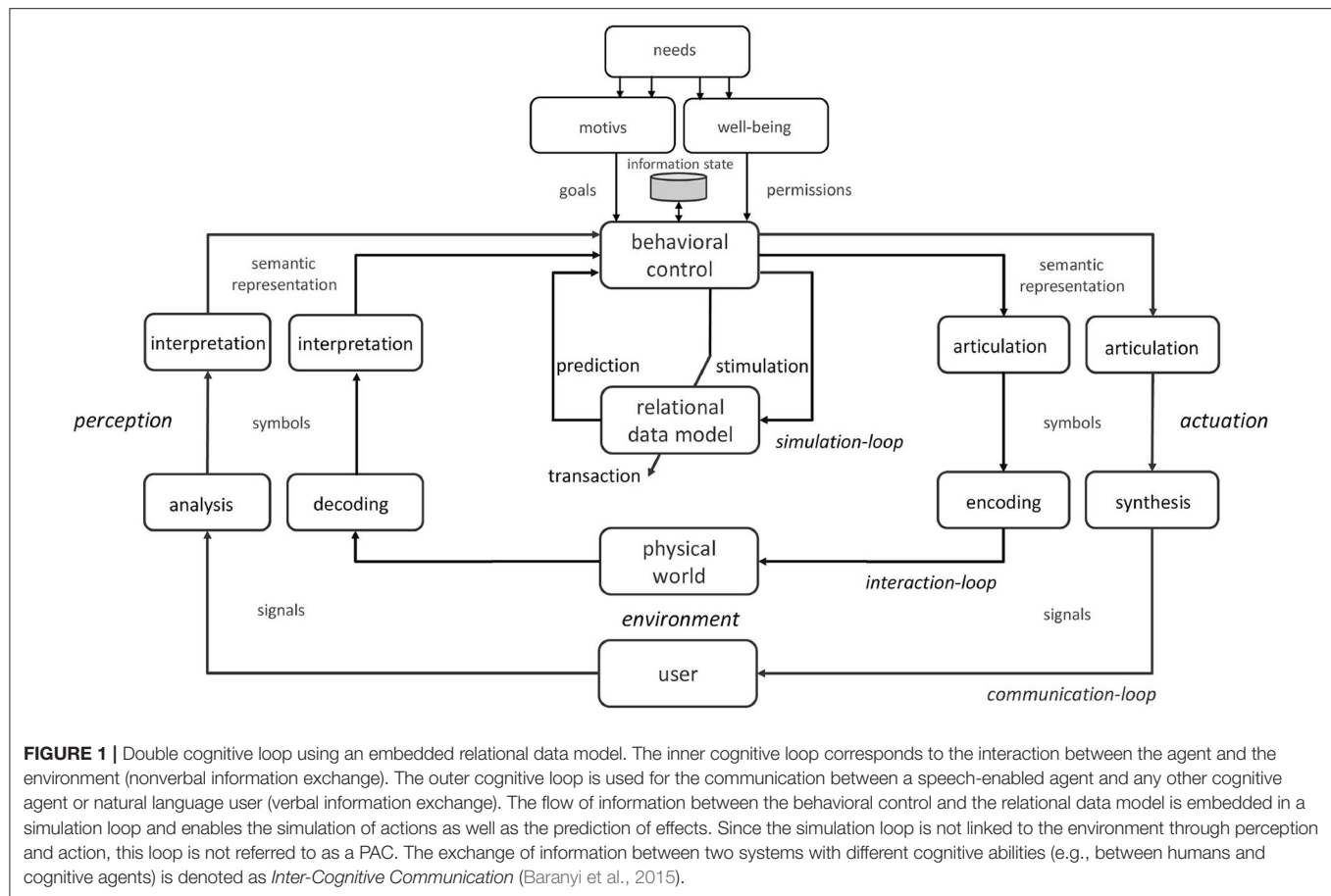
In our application scenario, the communications between cognitive agents and human users are related to a common physical environment that is modeled through a network of objects that are in static or dynamic relationships with one another. Such objects can be described and distinguished by a set of attributes and its admissible values. Objects, attributes and relationships between objects are the information of interest to which both nonverbal interaction and verbal communication refer. However, in order to ensure veridical behavior, the representational data structures of the environment model must be compared with those that have resulted from communication and interaction. Only by using an appropriate comparison mechanism the agent can understand what is actually going on in reality and what the respective observations mean. It largely depends on this ability whether the right actions are selected to achieve the agent's goals. In case of interaction, the comparison must be made between the representational data of the sensor information and the content of the agents knowledge base. In communication, on the other hand, it is necessary that the representational data structures of the user's utterances (semantics) are referenced to the shared environmental context (Hausser, 2014). The agent can receive this information using the interaction loop. Note that this comparison may result in an adaptation of the knowledge base.

Research in computational linguistics has demonstrated that quite different grammar formalisms, such as tree-adjoining grammar (Joshi et al., 1975), multiple context-free grammar (Seki et al., 1991), range concatenation grammar (Boullier, 2005), and minimalist grammar (Stabler, 1997; Stabler and Keenan, 2003) converge toward universal description models (Joshi et al., 1990; Michaelis, 2001; Stabler, 2011a; Kuhlmann et al., 2015). Minimalist grammar has been developed by Stabler (1997) to mathematically codify Chomsky's *Minimalist Program* (Chomsky, 1995) in the generative grammar framework. A minimalist grammar consists of a mental lexicon storing linguistic signs as arrays of syntactic, phonetic and semantic features, on the one hand, and of two structure-building functions, called "merge" and "move," on the other hand. Furthermore, syntax and compositional semantics can be combined via the lambda calculus (Niyogi, 2001; Kobele, 2009), while MG parsing can be straightforwardly implemented through bottom-up (Harkema, 2001), top-down (Harkema, 2001; Mainguy, 2010; Stabler, 2011b), and in the meantime also by left-corner automata (Stanojević and Stabler, 2018).

One important property of MG is their effective learnability in the sense of Gold's formal learning theory (Gold, 1967). Specifically, MG can be acquired by positive examples (Bonato and Retoré, 2001; Kobele et al., 2002; Stabler et al., 2003) from linguistic dependence graphs (Nivre, 2003; Klein and Manning, 2004; Boston et al., 2010), which is consistent with

¹Based on the terms used in control theory, we represent the cyclical flow of information in a circular form and denote the cyclical flow as a loop.

²Note, that the opposite point of view is also of interest, in which the agent learns user behavior by analyzing dialogs.



psycholinguistic findings on early-child language acquisition (Gee, 1994; Pinker, 1995; Ellis, 2006; Tomasello, 2006; Diessel, 2013). However, learning through positive examples only, could easily lead to overgeneralization. According to Pinker (1995) this could substantially be avoided through reinforcement learning (Skinner, 2015; Sutton and Barto, 2018). Although there is only little psycholinguistic evidence for reinforcement learning in human language acquisition (Moerk, 1983; Sundberg et al., 1996), we outline a machine learning algorithm for the acquisition of an MG mental lexicon (beim Graben et al., 2020) of the syntax and semantics for English declarative sentences through reinforcement learning in this article. Our approach chosen here is inspired by similar work on linguistic reinforcement learning (Zettlemoyer and Collins, 2005; Kwiatkowski et al., 2012; Artzi and Zettlemoyer, 2013). Instead of looking at pure syntactic dependencies as Bonato and Retoré (2001), Kobele et al. (2002), and Stabler et al. (2003), our approach directly uses their underlying semantic dependencies for the simultaneous segmentation of syntax and semantics, in some analogy to van Zaanen (2001).

With this work we are pursuing two main goals. The first goal is to represent the received or released information about the properties of objects or the relationships between objects in the form of relational expressions and to store it in a knowledge model so that these expressions can be used for knowledge processing for the purpose of behavior control or

problem solving. In order to assign a meaning and a context to the symbolic structures (symbol grounding problem), the comparability between the expressions of the knowledge model and the expressions of the two cognitive loops (interaction and communication) must be guaranteed. We fulfill this requirement by uniformly describing the information transformation in both loops using the formalism of minimalist grammar and thus with linguistic means. To solve the symbol grounding problem, the relationship between the symbol structures and the actual facts is established through a model of the agent's environment. Such a model can be acquired on the principle of reinforcement learning. Since the information transformation is also based on a model—here in the form of the mental lexicon—the second aim of this work is to acquire the mental lexicon of an MG based on this learning principle.

The paper is organized as follows: First, we address the problem statement and the experimental environment. Subsequently, we summarize the necessary mathematical basics and the notations used. Then we come to the first focus of this work and describe the unified information transformation for interaction and communication using minimalist grammar and lambda calculus. We start with the simple case of interaction and explain the transformation mechanism using formal linguistic means and the assumption that the agent has already acquired a minimalist lexicon. Subsequently, the more demanding case of communication is discussed, where

the transformation mechanism has to convert natural language utterances into predicate logical expressions. In both cases we discuss language production (articulation) and language understanding (interpretation). The integration of the acquired representational data structures into the agent's knowledge base concludes the first main focus. The following section covers the second main focus and deals with the acquisition of minimalist lexicons through the method of reinforcement learning. The paper concludes with a summary and a discussion of the achieved results.

2. PROBLEM STATEMENT AND EXPERIMENTAL SETUP

We consider the well-known mouse-maze problem (Shannon, 1953; Wolff et al., 2015, 2018), where an artificial mouse lives in a simple $N \times M$ maze world that is given by a certain configuration of walls. For the mouse to survive, one or more target objects are located at some places in the maze. In our setup we are using the target object types cheese/carrot (C) and water (W) to satisfy the primary needs hunger and thirst. These object types are defined symbolically by the set $\mathcal{O} = \{C, W, NOB\}$ where NOB refers to no object at all. The agent is able to move around the maze and to perceive information about its environment via physical interaction. This is organized by the inner perception-action-cycle (interaction-loop) of **Figure 1** that allows the agent to navigate to these target objects. To this end, the agent needs to measure its current position (x, y) by two sensors, where $x \in X = \{1, \dots, N\}$ and $y \in Y = \{1, \dots, M\}$ apply. It also has to determine the presence or absence of target objects by an object classifier, which we consider as a single complex sensor. Then the current situation can be described on the basis of the measurement result for the current position and the result of object classification. Subsequently, the measurement information needs to be encoded as a string of symbols and has to be translated into a representational data structure saved in a knowledge base. This kind of knowledge ("situations") should be stored as the result of an exploration phase if no logical contradictions occur. A further kind of knowledge is the set of movements in the maze. These "movements" are based on permissible actions $a \in \mathcal{A}$, which initially correspond to the four geographic directions, north (N), south (S), west (W), and east (E), that are defined symbolically by the set $\mathcal{A} = \{N, S, W, E, NOP\}$ (NOP denoting no operation here). Each action starts at a position $z = (x, y)$ and ends at a position $z' = (x', y')$. For this purpose, the actuators associated to the x - or y - direction, respectively, can be incremented or decremented by one step. Hence, to establish the parameterization of the four geographic directions we define the sets $\Delta X = \Delta Y = \{-1, 0, 1\}$. Thus, the south action is parameterized, for example, by the ordered pair $(0, -1)$. Note that with the knowledge about "movements" the agents behavior can be described in the sense of "causality." From a technical point of view, the relationship between a cause $[z = ((x, y), a)]$ and an effect $[z' = (x', y')]$ can be expressed, for example, by the transition equation of a finite state automaton. To implement "movement" instructions the reverse flow of the sensory information must be realized. That is, in order to be

able to control the agent's actuators, actions must be described as representative data structures and converted into a string of symbols. It follows, that physical interaction requires both, the agent's capability to interpret a linearly ordered time series of symbolized measurement results as a (partially ordered) semantic representations and to transform semantic representations into a linear sequence of actuator instructions during articulation.

The ability to communicate with natural language users is another demand that a cognitive agent should meet. To this end, the agent should first of all be speech-enabled. Communication is organized by the outer perception-action-cycle (communication-loop) shown in **Figure 1** which is mainly characterized by the articulation and interpretation of speech signals. Therefore, we essentially need the very same capabilities for the transformation of linearly ordered symbolic messages (the "scores" of communication) into partially ordered semantic representations. For this reason, we devise a *bidirectional* utterance meaning transducer to encode and decode the meaning of symbolic messages. The encoded meaning corresponds to the representational data structure of utterances and can be saved in the agent's knowledge base.

In order to avoid logical conflicts between the results of the interaction- and communication loop, it is also necessary that the representational data structures generated by these loops must be comparable. Hence, in our approach non-verbal physical interaction and verbal communication are uniformly modeled using linguistic description means. This approach can be also supported by two salient arguments borrowed from Hausser (2014) that have already been presented by Römer et al. (2019): (1) "Without a carefully built physical grounding any symbolic representation will be mismatched to the sensors and actuators. These groundings provide the constraints on symbols necessary for them to be truly useful." (2) "The analysis of nonverbal cognition is needed in order to be able to plausibly explain the phylogenetic and ontogenetic development of language from earlier stages of evolution without language." Further, according to the "Physical Symbol Systems Hypothesis" (PSSH) all cognitive processes can be described as the transformation of symbol structures (Newell and Simon, 1976). This transformation process obeys the composition principle ("infinite use of finite means") and aims to recast incoming sensor information into logically processable knowledge, which is saved in a knowledge model. Notably, we assume that the transformation from signal to symbol space can be solved by a transduction stage, proposed by Wolff et al. (2013). According to the PSSH, this low-level transduction stage is the crucial processing step for recognition, designation and arrangement of the observed sensor events through symbol sequences in terms of a formal or natural language.

The use of linguistic description means includes the use of a suitable grammar formalism that is based on a mental lexicon as part of the agent's knowledge base and on transformation rules for how to arrange symbols into linear sequences. It is the second aim of the present study, to suggest minimalist grammar and logical lambda calculus as a unifying framework to this end. While the user can be assumed to already have the linguistic resources, the agent must acquire them through learning. Thus, a further aim of our work is the description of an algorithm with

which a mental lexicon can be learned and dynamically updated through reinforcement learning. Starting point of our algorithm are *utterance meaning pairs* (Kwiatkowski et al., 2012; Wirsching and Lorenz, 2013; beim Graben et al., 2019b).

$$u = \langle e, \sigma \rangle, \quad (1)$$

where $e \in E$ is the spoken or written utterance while $\sigma \in \Sigma$ is a logical term, expressed by means of predicate logic and the (untyped) lambda calculus (Church, 1936) denoting the *semantics* of the utterance. We assume that UMPs are continuously delivered by a “teacher” to the agent during language acquisition.

As an example, consider the simple UMP

$$u = \langle \text{the mouse eats cheese}, \text{eat}(\text{cheese})(\text{mouse}) \rangle. \quad (2)$$

In the sequel we use typewriter font to emphasize that utterances are regarded plainly as symbolic tokens without any intended meaning in the first place. This applies even to the “semantic” representation in terms of first order predicate logic where we use the Schönfinkel-Curry (Schönfinkel, 1924; Lohnstein, 2011) notation here. Therefore, the expression above $\text{eat}(\text{cheese})(\text{mouse})$ indicates that eat is a binary predicate, fetching first its direct object cheese to form a unary predicate, $\text{eat}(\text{cheese})$, that then takes its subject mouse in the second step to build the proposition of the utterance (2).

3. PRELIMINARIES

In this section we summarize some fundamental concepts needed for system modeling and information transformation: state space representation and formal languages. We also briefly refer to the set-theoretical connection between simple predicate logic expressions (without quantifiers) and semantic representations based on relational schemes. Subsequently, we explain the concept of minimalist grammar and its suitability for a transformation mechanism that can be used to mediate between linearly ordered symbolic perception/action sequences and semi- or disordered semantic representations. The expressiveness and flexibility of natural language is mainly due to the compositional principle, according to which new semantic terms can be formed from a few elementary terms. In order to anchor this principle in the transformation mechanism, we need the approach of the lambda calculus and the representation of n -ary functions using the Schönfinkel-Curry notation.

3.1. Basic Concepts

State Space Representation. To describe the system behavior in terms of cause and effect, we distinguish states $z \in \mathcal{Z}$, actions $a \in \mathcal{A}$ and outputs³ $o \in \mathcal{O}$. A behavioral relation is then given by

$$R_V \subseteq \mathcal{Z} \times \mathcal{A} \times \mathcal{Z} \times \mathcal{O}, \quad (3)$$

³To avoid confusion in terms of the set \mathcal{O} , we would like to point out that the outputs of the state space model in this work correspond to the specified object types of the mouse-maze-application.

which determines the system dynamics. To decide, whether a tuple (z, a, z', o) belongs to this relation or not, the characteristic function $f_{R_V}: \mathcal{Z} \times \mathcal{A} \times \mathcal{Z} \times \mathcal{O} \rightarrow \{0, 1\}$ can be applied. Further, this relation corresponds to a network of causal relations. Due to causality, a recursive calculation rule divided into a system equation and an output equation is given, which allows forecasting the system's behavior

$$\begin{aligned} z' &= G(z, a) & \text{with } G: \mathcal{Z} \times \mathcal{A} &\rightarrow \mathcal{Z}, \\ o &= H(z, a) & \text{with } H: \mathcal{Z} \times \mathcal{A} &\rightarrow \mathcal{O}. \end{aligned} \quad (4)$$

Because in our setting the system response $o \in \mathcal{O}$ depends exclusively on the current state $z \in \mathcal{Z}$, we get a simplification of the output equation $o = H(z)$ that corresponds to the idea of the Moore automaton.

Formal Language. In order to specify any technical communication between source and sink, we have to arrange an alphabet $\Sigma_A = \{a, b, \dots\}$ and to establish a language $\mathcal{L} \subseteq \Sigma_A^*$ (the symbol $*$ is called Kleene star). Words or sentences are then described by an ordered sequence $\mathbf{s} = (s_1, s_2, \dots, s_k) \in \mathcal{L}$, where $s \in \Sigma_A$. To decide whether a word s belongs to the language \mathcal{L} we can devise a grammar $G = (\mathcal{N}, \mathcal{T}, \mathcal{P}, S)$. It is specified by the sets \mathcal{N} of nonterminals, \mathcal{T} of terminals and \mathcal{P} of production rules as well as a start symbol S . The grammatical rules determine the arrangement of the symbols within a symbol sequence. Based on these rules permissible sentences of a language can be produced or derived. In contrast to sentence production, sentence analysis reveals the underlying grammatical structure. This analysis method is known as parsing and will be exploited below.

3.2. Predicate Logic Expressions

A linguistic expression that articulates the characteristics of an object or a relation between objects is called a predicate (Jungclaussen, 2001). It does not always have to be about definite objects. In the case of indefinite objects, so-called individual variables are used. Usually the letters P or Q are used as symbols for predicate names. As arguments for predicates we use either individual constants, individual variables or function values. For the sake of simplicity, we limit ourselves to individual constants and individual variables here. An n -ary predicate is then denoted by the expression $Q(a_1, a_2, \dots, a_n)$, where a are values of the corresponding set A (e.g., admissible values from the domain of an attribute A). According to set theory the predicate $Q(a_1, a_2, \dots, a_n)$ can be interpreted as relation $R_Q \subseteq A_1 \times A_2 \times \dots \times A_n$ and can be understood as a binary classifier. In this case the classification task refers to finding in the set of ordered tuples of individuals or objects the class $[Q]$ of those individuals for which the predicate is true. This class is called the denotation of Q . The following notation is used for this:

$$[Q] = \{(a_1, \dots, a_n) \in A_1 \times \dots \times A_n \mid Q(a_1, \dots, a_n) \text{ is true} \} \quad (5)$$

The mathematical description of a predicate can also be done using the corresponding characteristic function: $f_{[Q]}: A_1 \times A_2 \times \dots \times A_n \rightarrow \{0, 1\}$. This is particularly required for the formalism of composing semantic representations.

3.3. Semantic Representations

In semantics we distinguish between objects that are described by attributes and relations between such objects (Chen, 1975). To represent semantics we will focus on feature-value pairs (FVP), which have a flat structure and correspond to tuples of database relations. The notation used here was taken from Lausen (2005). We start from a universe $\mathcal{U} = \{A_1, A_2, \dots, A_m\}$ which comprises a finite set of attributes. Further, we have a set of domains $\mathcal{D} = \{D_1, D_2, \dots, D_m\}$ and a mapping $\text{dom} : \mathcal{U} \rightarrow \mathcal{D}$. The values of the different attributes are elements of the associated domains. Based on these definitions we can specify types of objects or relations between objects by a set of attributes $\mathcal{X} = \{A_1, A_2, \dots, A_n\} \subset \mathcal{U}$. Objects are defined as mappings, which are called tuples:

$$\tau : \{A_1, A_2, \dots, A_n\} \rightarrow \bigcup_{i=1}^n \text{dom}(A_i), n \leq m. \quad (6)$$

A set of such tuples corresponds to a set of distinguishable objects. Note that in database notation the elements of tuples are not ordered. This corresponds to the idea that the order of attributes is not relevant for describing the type of any object (Lausen, 2005). An object or relation type is defined by a relation scheme:

$$R(\mathcal{X}) = (A_1 : \text{dom}(A_1), A_2 : \text{dom}(A_2), \dots, A_n : \text{dom}(A_n)). \quad (7)$$

In database representations a scheme is associated to the head of a relational data table. The entries of a table correspond to a set of tuples. This set of tuples corresponds to a relation, which is given by:

$$R_{\mathcal{X}} \subseteq \text{Tup}(X) := \{\tau \mid \tau : X \rightarrow \text{dom}(\mathcal{X})\}, \quad (8)$$

where $\text{Tup}(\mathcal{X})$ is the set of all possible tuples of a relation scheme $R(\mathcal{X})$. The set of all relations over this scheme is denoted as $\text{Rel}(\mathcal{X})$.

3.4. Minimalist Grammars

Minimalist grammars (MG) were introduced to describe natural languages (Stabler, 1997). For nonverbal interaction it is crucial that MG provide a translation mechanism between linearly ordered perceptions/actions and semi- or disordered semantic representations. Based on such a grammar, knowledge can be formally represented and stored consistently in a relational data model. Following Kracht (2003), we regard a linguistic sign as an ordered triple

$$q = \langle e, t, \sigma \rangle \quad (9)$$

with exponent $e \in E$, semantics $\sigma \in \Sigma$ and a syntactic type $t \in T$ that we encode by means of MG in its chain representation (Stabler and Keenan, 2003). The type controls the generation of the syntactic and semantic structure of an utterance. An MG consists of a data base, the mental lexicon, containing signs as arrays of syntactic, phonetic and semantic *features*, and of two structure-generating functions, called “merge” and “move.” For our purposes, it is sufficient to point out some syntactic features: Selectors (e.g., “=S”) are required to search for syntactic types of the same category (e.g., “S”). Licensors (e.g., “+k”) and licensees (e.g., “-k”) are required to control the symbol

order at the surface. The symbol “::” indicates *simple, lexical* categories while “:” denotes *complex, derived* categories (“.” is just a placeholder for one of these signs). A sequence of signs “q” is called a *minimalist expression*. For further details we refer to beim Graben et al. (2019b). We just repeat the two kinds of structure-building rules here. The MG function “merge” is defined through inference schemes

$$\begin{array}{ll} \frac{\langle e_1, ::=ft, \sigma_1 \rangle \quad \langle e_2, f, \sigma_2 \rangle q}{\langle e_1 e_2, :t, \sigma_1 \sigma_2 \rangle q} & \text{merge-1,} \\ \frac{\langle e_1, ::=ft, \sigma_1 \rangle q_1 \quad \langle e_2, f, \sigma_2 \rangle q_2}{\langle e_2 e_1, :t, \sigma_1 \sigma_2 \rangle q_1 q_2} & \text{merge-2,} \\ \frac{\langle e_1, ::=ft, \sigma_1 \rangle q_1 \quad \langle e_2, f, \sigma_2 \rangle q_2}{\langle e_1, :t, \sigma_1 \rangle q_1 \langle e_2, :t, \sigma_2 \rangle q_2} & \text{merge-3.} \end{array} \quad (10)$$

Correspondingly, “move” is given through,

$$\begin{array}{ll} \frac{\langle e_1, ::+ft, \sigma_1 \rangle q_1 \langle e_2, :-f, \sigma_2 \rangle q_2}{\langle e_2 e_1, :t, \sigma_1 \sigma_2 \rangle q_1 q_2} & \text{move-1,} \\ \frac{\langle e_1, ::+ft, \sigma_1 \rangle q_1 \langle e_2, :-f, \sigma_2 \rangle q_2}{\langle e_1, :t, \sigma_1 \rangle q_1 \langle e_2, :t, \sigma_2 \rangle q_2} & \text{move-2.} \end{array} \quad (11)$$

3.5. Schönfinkel-Curry Notation and Lambda Calculus

Schönfinkel-Curry Notation. Let us consider a binary function

$$f : A \times B \rightarrow Z, \quad (12)$$

where its domain is defined by the cartesian product over two sets A and B and its value range is given by the set Z . Such a binary function can be decomposed into two unary functions, which are calculated in two steps

$$F : A \rightarrow (B \rightarrow Z). \quad (13)$$

First, with the assignment of the first argument $a \in A$, a mapping $F(a) : B \rightarrow Z$ is selected. Secondly, the function value $F(a)(b) = f(a, b) = z$ is calculated when the second argument $b \in B$ is assigned. This principle can be generalized to n -ary functions

$$f : A_1 \times A_2 \times \dots \times A_n \rightarrow A_{n+1}. \quad (14)$$

Whereby a one-to-one relationship between n -ary functions and unary functions of n -th order is established

$$F : A_1 \rightarrow (A_2 \rightarrow (A_3 \dots (A_n \rightarrow A_{n+1}) \dots)). \quad (15)$$

In computational linguistics this representation is used in terms of the description of relations by their characteristic function. In this case an n -tuple $(a_1, a_2, \dots, a_n) \in A_1 \times A_2 \times \dots \times A_n$ is mapped to the elements of the binary set $\{0, 1\}$. However, the representation in computational linguistics is in a slightly modified form, because the arguments appear in the reverse order $F(a_n)(a_{n-1}) \dots (a_1)$.

Lambda calculus is a mathematical formalism developed by Church in the 1930s “to model the mathematical notion of substitution of values for bound variables” according to

Wegner (2003). Although the original application was in the area of computability (cf. Church, 1936) the substitution of parts of a term with other terms is often the central notion when lambda calculus is used. This is also true in our case and we have to clarify the concepts first; namely *variable*, *bound* and *free*, *term*, and *substitution*.

To be applicable to any universe of discourse a prerequisite of lambda calculus is “an enumerably infinite set of symbols” (Church, 1936) which can be used as variables. However, for usage in a specific domain, a finite set is sufficient. Since we aim at terms from first order predicate logic, treating them with the operations from lambda calculus, all their predicates P and individuals I need to be in the set of variables. Additionally, we will use the symbols $I_V := \{x, y, \dots\}$ as variables for individuals and $T_V := \{P, Q, \dots\}$ as variables for (parts of) logical terms. The set $V := P \cup I \cup I_V \cup T_V$ is thus used as the set of variables. Note, that the distinction made by I_V and T_V is not on the level of lambda calculus but rather a meta-theoretical clue for the reader.

The following definitions hold for lambda calculus in general and hence we simply use “normal” typeface letters for variables. The term algebra of lambda calculus is inductively defined as follows. *i)* Every variable $v \in V$ is a term and v is a *free* variable in the term v ; specifically, also every well-formed formula of predicate logic is a term. *ii)* Given a term T and a variable $v \in V$ which is free in T , the expression $\lambda v.T$ is also a term and the variable v is now *bound* in $\lambda v.T$. Every other variable in T different from v is free resp. bound in $\lambda v.T$ if it is free resp. bound in T . *iii)* Given two terms T and U , the expression $T(U)$ is also a term and every variable which is free resp. bound in T or U is free resp. bound in $T(U)$. Such a term is often referred to as *operator-operand combination* (Wegner, 2003) or *functional application* (Lohnstein, 2011). For disambiguation we also allow parentheses around terms. The introduced syntax differs from the original one where additionally braces and brackets are used to mark the different types of terms (cf. Church, 1936). Sometimes, $T(U)$ is also written as (TU) and the dot between λ and the variable is left out (cf. Wegner, 2003).

For a given variable $v \in V$ and two terms T and U the operation of *substitution* is $T[v \leftarrow U]$ [originally written as $S_v^U T$ in Church (1936) and sometimes without the right bar, i.e. as in Wegner (2003)] and stands for the result of substituting U for all instances of v in T .

Church defined three conversions based on substitution.

- *Renaming* bound variables by replacing any part $\lambda v.T$ of a term by $\lambda w.T[v \leftarrow w]$ when the variable w does not occur in the term T .
- *Lambda application* by replacing any part $\lambda v.T(U)$ of a term by $T[v \leftarrow U]$, when the bound variables in T are distinct both from v and the free variables in U .
- *Lambda abstraction* by replacing any part $T[v \leftarrow U]$ of a term by $\lambda v.T(U)$, when the bound variables in T are distinct both from v and the free variables in U .

The first conversion simply states that names of bound variables have no particular meaning on their own. The second and third conversions are of special interest to our aims. Lambda

application allows the composition of logical terms out of predicates, individuals and other logical terms while lambda abstraction allows the creation of templates of logical terms.

Applied to our example (2), we have the sign

$$q = (\text{the mouse eats cheese}, :c, \text{eat(cheese)(mouse)}) \quad (16)$$

where the now appearing MG type $:c$ indicates that the sign is complex (not lexical) and a complementizer phrase of type c . Its compositional semantics (Lohnstein, 2011) can be described by the terms $\lambda P.\lambda x.P(x)$ and $\lambda x.\lambda P.P(x)$, the predicate *eat* and the individuals *cheese* and *mouse*. Consider the term $\lambda P.\lambda x.P(x)(\text{eat})(\text{cheese})$. This is converted by two successive lambda applications via $\lambda x.\text{eat}(x)(\text{cheese})$ into the logical term eat(cheese) . It is also possible to rearrange parts of the term in a different way. Consider now the term $\lambda x.\lambda P.P(x)$, the logical term eat(cheese) and the individual *mouse*. Then the term $\lambda x.\lambda P.P(x)(\text{mouse})(\text{eat(cheese)})$ is converted by two successive lambda applications into the logical term $\text{eat(cheese)(mouse)}$. Thus, logical terms can be composed through lambda application.

Moreover, given the logical term $\text{eat(cheese)(mouse)}$ two successive lambda abstractions yield the term $\lambda x.\lambda y.\text{eat}(x)(y)$, leaving out the operand parts. In that way, templates of logical terms are created where different individuals can be inserted for term evaluation. Both processes are crucial for our utterance-meaning transducer and machine language acquisition algorithms below.

4. PHYSICAL INTERACTION

In order to explain the translation process for the interaction, we first rely on a regular grammar, which only comprises two rules: $S \rightarrow eS$ and $S \rightarrow e$ with $e \in \mathcal{T}$ and $S \in \mathcal{N}$. To better understand the formalism we use the following indexed derivation scheme: $S_0 \rightarrow e_0 S_1, S_1 \rightarrow e_1 S_2, S_2 \rightarrow e_2 S_3, \dots, S_n \rightarrow e_n S_{n+1}, S_{n+1} \rightarrow e_{n+1}$. The inner words e_1, e_2, \dots, e_n correspond to the values of attributes that are specified in a relational scheme $R(\mathcal{X})$. The embracing words e_0 and e_{n+1} correspond to the symbols $\langle \text{start} \rangle$ and $\langle \text{end} \rangle$, respectively, which indicate the beginning and the end of a sentence. The database semantics is made up of sets, the elements of which are assignments of values to attributes $\tau: A \rightarrow e \in \text{dom}(A)$. To apply the translation formalism based on minimalist grammars we substitute semantic terms σ with $\sigma(e)$. Then we define the semantic concatenation as a set operation: $\sigma_1 \sigma_2 := \sigma(e_1) \cup \sigma(e_2) := \{\sigma(e_1)\} \cup \{\sigma(e_2)\}$.

Minimalist lexicon. The MG formalism works with linguistic signs that are saved in a minimalist lexicon. In case of interaction three types of linguistic signs that are saved in a minimalist lexicon (see Table 1) are required.

The semantics of these sign types are represented by λ -terms using the set-theoretical union operator in the Schönfinkel-Curry-notation $\cup(\sigma(e_2))(\sigma(e_1)) := \cup(\sigma(e_1), \sigma(e_2)) = \sigma(e_1) \cup \sigma(e_2)$. According to the following formalism in each iteration step exactly one mapping is added to the semantic expression generated so far. In λ -calculus the composition is achieved

TABLE 1 | Minimalist lexicon for physical interaction.

$\langle \text{start} \rangle, :: S_0 \ k, (\sigma(e_0))$
 $\langle \text{end} \rangle, :: S_n + k \ S_{n+1}, \lambda a. \cup (\sigma(e_{n+1}))(a)$
 $\langle e_i, :: S_{i-1} + k \ S_i, \lambda a. \cup (\sigma(e_i))(a) \rangle$

The symbols $\langle \text{start} \rangle = e_0$ and $\langle \text{end} \rangle = e_{n+1}$ encode the beginning and the end of a sequence of symbols. The embedded entries e_i correspond to sensory events. The index i with $0 < i < n + 1$ is related to the attribute of the relational scheme. The semantics of e_i is given by $\sigma(e_i)$, whereby $\sigma(e_0) = \sigma(e_{n+1}) = \emptyset$ is applied.

through consecutive λ -applications, where the position for the argument a is alternately opened and closed. In this way, mappings are added one after the other to a disordered set, which results to the entries of a relational scheme.

4.1. Interpretation

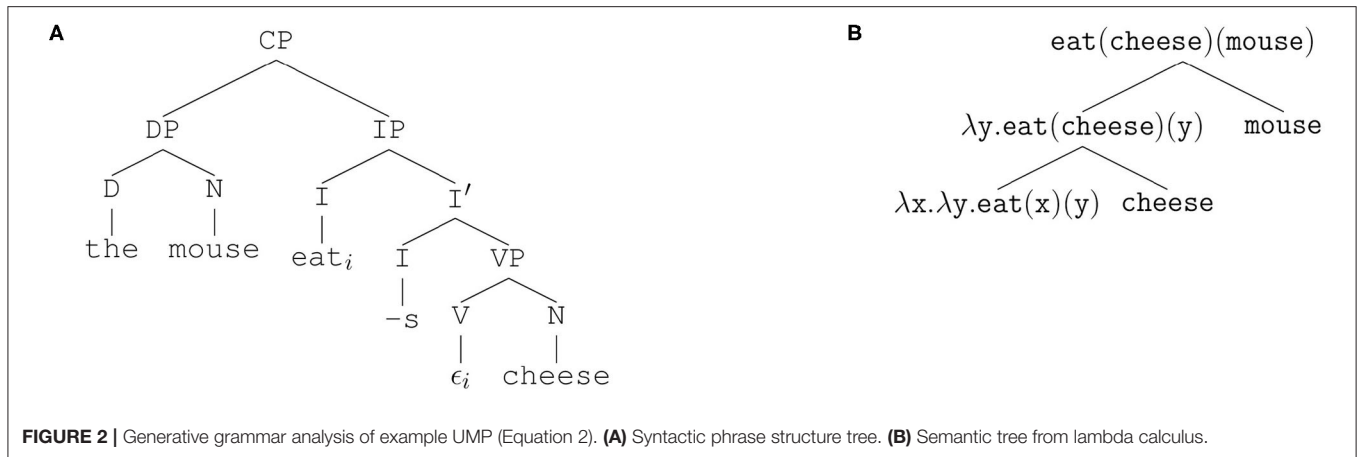
During the interpretation process, a sequence of symbols is converted into a set of mappings that describe a situation. Such

Permissible sentences are elements of the observation language \mathcal{L}_{obs} . The configuration of these sentences is subject to a fixed agreement. For this purpose we assign a role to every exponent index and use a bijective mapping $\beta: \mathcal{R} \rightarrow \mathcal{X}$, which maps each role $r \in \mathcal{R}$ to the attributes $A \in \mathcal{X}$ of the relational schemes $R(\mathcal{X})$. The semantics of an exponent e_i with $0 < i < n + 1$ is then defined by $\sigma(e_i): \beta(r_i) \mapsto \omega_i \in \text{dom}(\beta(r_i))$ and $\sigma(e_0) = \sigma(e_{n+1}) = \emptyset$. Based on this role-dependent semantics we obtain with the MG formalism a linguistic structure in which the syntactic and semantic structure are built up in parallel. At its core, a *key-lock principle* is applied, in which a selector (e.g. “=”) always requests a syntactic type of the same category (e.g. “S”) and licensors and licensees are used to control the order of the symbols on the surface. Note that this principle is not restricted to natural languages. Now we come to the description of the formalism using the λ -calculus. First, the symbol sequence $\mathbf{s} \in \mathcal{L}_{obs}$ is converted into a *sequence of linguistic signs* (taken from the minimalist lexicon), which the formalism processes step

$$\begin{array}{l}
 \text{Start:} \quad \frac{\langle e_1, :: S_0 + k S_1 - k, \lambda a. \cup (\sigma(e_1))(a) \rangle \quad \langle e_0, :: S_0 - k, \sigma(e_0) \rangle}{\langle e_1, :: +k S_1 - k, \lambda a. \cup (\sigma(e_1))(a) \rangle \quad \langle e_0, :: -k, \sigma(e_0) \rangle} \text{merge-3} \\
 \quad \frac{\langle e_1, :: +k S_1 - k, \lambda a. \cup (\sigma(e_1))(a) \rangle \quad \langle e_0, :: -k, \sigma(e_0) \rangle}{\langle e_0 e_1, :: S_1 - k, \cup (\sigma(e_1))(\sigma(e_0)) \rangle} \text{move-1.} \\
 \\
 \text{Iterations:} \quad \frac{\langle e_2, :: S_1 + k S_2 - k, \lambda a. \cup (\sigma(e_2))(a) \rangle \quad \langle e_0 e_1, :: S_1 - k, \bigcup_{i=0}^1 \sigma(e_i) \rangle}{\langle e_2, :: +k S_2 - k, \lambda a. \cup (\sigma(e_2))(a) \rangle \quad \langle e_0 e_1, :: -k, \bigcup_{i=0}^1 \sigma(e_i) \rangle} \text{merge-3} \\
 \quad \frac{\langle e_2, :: +k S_2 - k, \lambda a. \cup (\sigma(e_2))(a) \rangle \quad \langle e_0 e_1, :: -k, \bigcup_{i=0}^1 \sigma(e_i) \rangle}{\langle e_0 e_1 e_2, :: S_2 - k, \bigcup_{i=0}^2 \sigma(e_i) \rangle} \text{move-1.} \\
 \\
 \quad \frac{\langle e_3, :: S_2 + k S_3 - k, \lambda a. \cup (\sigma(e_3))(a) \rangle \quad \langle e_0 e_1 e_2, :: S_2 - k, \bigcup_{i=0}^2 \sigma(e_i) \rangle}{\langle e_3, :: +k S_3 - k, \lambda a. \cup (\sigma(e_3))(a) \rangle \quad \langle e_0 e_1 e_2, :: -k, \bigcup_{i=0}^2 \sigma(e_i) \rangle} \text{merge-3} \\
 \quad \frac{\langle e_3, :: +k S_3 - k, \lambda a. \cup (\sigma(e_3))(a) \rangle \quad \langle e_0 e_1 e_2, :: -k, \bigcup_{i=0}^2 \sigma(e_i) \rangle}{\langle e_0 e_1 e_2 e_3, :: S_3 - k, \bigcup_{i=0}^3 \sigma(e_i) \rangle} \text{move-1.} \\
 \\
 \text{End:} \quad \frac{\langle e_4, :: S_3 + k S_4, \lambda a. \cup (\sigma(e_4))(a) \rangle \quad \langle e_0 e_1 e_2 e_3, :: S_3 - k, \bigcup_{i=0}^3 \sigma(e_i) \rangle}{\langle e_4, :: +k S_4, \lambda a. \cup (\sigma(e_4))(a) \rangle \quad \langle e_0 e_1 e_2 e_3, :: -k, \bigcup_{i=0}^3 \sigma(e_i) \rangle} \text{merge-3} \\
 \quad \frac{\langle e_4, :: +k S_4, \lambda a. \cup (\sigma(e_4))(a) \rangle \quad \langle e_0 e_1 e_2 e_3, :: -k, \bigcup_{i=0}^3 \sigma(e_i) \rangle}{\langle e_0 e_1 e_2 e_3 e_4, :: S_4, \bigcup_{i=0}^4 \sigma(e_i) \rangle} \text{move-1.}
 \end{array}$$

situations are encoded in the form: $\mathbf{s} = (e_0, e_1, e_2, e_3, e_4)$ which corresponds in the mouse-maze scenario to messages $\mathbf{s} = (\langle \text{start} \rangle, x, y, o, \langle \text{end} \rangle)$. That is, the exponents contain the (x, y) -coordinates and the name of an object type $o \in \mathcal{O}$.

by step. During processing, the formalism alternates between the rules merge-3 and move-1. After processing the start type, the iterative processing of the mapping types takes place until the end type is reached. To calculate the semantics of \mathbf{s} , the



argument position for the new set element to be integrated is opened and closed again by consecutive λ -applications after the current feature-value pair has been added. Finally, the formalism replies with a tuple $\tau_{obs} \in Tup(Situation)$.

The last compositional step generates the linguistic sign $\langle e_0 e_1 e_2 e_3 e_4, :S_4, \bigcup_{i=0}^4 \sigma(e_i) \rangle$. The associated semantics corresponds to the structure of tuples, as defined for relational schemes according to Equation (6). If this formalism is applied to the relational scheme $R(Situation)$ (for relation scheme definitions see section 5.4) it results to the semantic representation of the observed situation

$$\tau_{obs} = \bigcup_{i=0}^{n+1} \sigma(e_i) = \{X \rightarrow x \in \text{dom}(X), Y \rightarrow y \in \text{dom}(Y), \mathcal{O} \rightarrow o \in \text{dom}(\mathcal{O})\}. \quad (17)$$

The observation tuple τ_{obs} contains the set of disordered feature-value-mappings, which corresponds to the translation result of the interpretation. Based on this tuple the next action is selected and a further perception-action-cycle is initiated.

4.2. Articulation

After the action decision by the behavior control (see **Figure 1**), the MG formalism must be applied to a semantic representation for the action in question. To this end, the formalism is fed with an action tuple $\tau_{act} \in Tup(Action)$. Each action tuple is initially given in the form.

$$\tau_{act} = \{\Delta X \rightarrow \Delta x \in \text{dom}(\Delta X), \Delta Y \rightarrow \Delta y \in \text{dom}(\Delta Y)\}. \quad (18)$$

In order to generate an actuator instruction $\mathbf{a} = (e_0, e_1, e_2, e_3)$ or $\mathbf{a} = (\langle \text{start} \rangle, \Delta x, \Delta y, \langle \text{end} \rangle)$, respectively, the feature value pairs of the relational scheme $R(Action)$ must be mapped to the associated linguistic signs. These signs are stored in the linguistic lexicon of the articulation (analogous to the interpretation). In this case too, the formalism alternately switches between the rules merge-3 and move-1. The difference to the interpretation is that the formalism is now fed with a *set of linguistic signs* and responds with a sequence of symbols. In our mouse-maze scenario this set comprises the

following elements: $\langle e_0, ::S_0-k, \emptyset \rangle$, $\langle e_3, ::S_2+kS_3, \emptyset \rangle$ as well as $\langle e_1, ::S_0+kS_1-k, \sigma(e_1) \rangle$ and $\langle e_2, ::S_1+kS_2-k, \sigma(e_2) \rangle$. Applying the production rules of a regular grammar, the derivation would result to the following rule sequence: $S_0 \rightarrow e_0 S_1$, $S_1 \rightarrow e_1 S_2$, $S_2 \rightarrow e_2 S_3$ and $S_3 \rightarrow e_3$. This results in $S_0 \rightarrow e_0 e_1 e_2 e_3$, in which the root of the derivation tree S_0 comprises the whole action sequence.

The MG-formalism preserves this order through the selectors of the associated exponents: After the formalism is starting with the sign $\langle e_0, ::S_0-k, \emptyset \rangle$ the exponent e_1 is required by the selector “ $=S_0$ ” of the sign $\langle e_1, ::S_0+kS_1-k, \sigma(e_1) \rangle$. Next, the exponent e_2 is required by the selector “ $=S_1$ ” of the sign $\langle e_2, ::S_1+kS_2-k, \sigma(e_2) \rangle$. Finally, the exponent e_3 is required by the selector “ $=S_2$ ” of the remaining sign in the articulation set $\langle e_3, ::S_2+kS_3, \emptyset \rangle$. The complete concatenated string $\mathbf{a} = (e_0, e_1, e_2, e_3) = (\langle \text{start} \rangle, \Delta x, \Delta y, \langle \text{end} \rangle)$ is inferred then by the last move-1 rule and corresponds to the action sequence that is subsequently transmitted to the agent’s actuators.

5. LINGUISTIC COMMUNICATION

As in the case of physical interaction, we also describe linguistic communication through a minimalist grammar. However, while interaction suffices with a minimalist implementation of regular grammars, exploiting only merge-3 and move-1 operations for the processing of linear time series of observation and actuation, natural language processing requires the full complexity of the MG formalism.

5.1. Utterance-Meaning Transducer (UMT)

In this section we propose a bidirectional *Utterance-Meaning-Transducer* (UMT) for both speech understanding and speech production by means of MG. For further illustrating the rules (10–11) and their applicability, let us stick with the example UMP (2) given in Sect. 2. Its syntactic analysis in terms of generative grammar (Haegeman, 1994) yields the (simplified) phrase structure tree in **Figure 2A**⁴.

⁴For the sake of simplicity we refrain from presenting full-fledged X-bar hierarchies (Haegeman, 1994).

The syntactic categories in **Figure 2A** are the *maximal projections* CP (complementizer phrase), IP (inflection phrase), VP (verbal phrase), and DP (determiner phrase). Furthermore, there are the intermediary node I' and the *heads* I (inflection), D (determiner), V (verb), and N (noun), corresponding to $t, d, v,$ and n in MG, respectively. Note that inflection is lexically realized only by the present tense suffix $-s$. Moreover, the verb *eat* has been moved out of its base-generated position leaving the empty string ϵ there. Movement is indicated by co-indexing with i .

Correspondingly, we present a simple semantic analysis in **Figure 2B** using the notation from Sect. 3.5 together with the lambda calculus of the binary predicate in its Schönfinkel-Curry notation (Schönfinkel, 1924; Lohnstein, 2011).

Guided by the linguistic analyses in **Figure 2**, an expert could construe a minimalist lexicon as given in **Table 2** by hand (Stabler and Keenan, 2003).

Semantics is given by the binary predicate $\text{eat}(x)(y)$ whose argument variables are bound by two lambda expressions $\lambda x. \lambda y$. Moreover, we have an inflection suffix $-s$ for present tense in third person singular, taking a predicate (pred) as complement, then triggering firstly inflection movement $+f$ and secondly case assignment $+k$, whose type is tense (t). Finally, there are two entries that are phonetically not realized. The first one selects a verbal phrase $=v$ and assigns case $+k$ afterwards; then, it selects a determiner phrase $=d$ as subject and has its own type predicate pred ; additionally, we prescribe an intertwiner of two abstract lambda expressions Q, P as its semantics. The last entry provides a simple type conversion from tense t to complementizer c in order to arrive at a well-formed sentence with start symbol c .

Using the lexicon (**Table 2**), the sign (16) is obtained by the minimalist derivation (19).

$$\frac{\langle \text{the}, ::=n \ d \ -k, \epsilon \rangle \quad \langle \text{mouse}, ::=n, \text{mouse} \rangle}{\langle \text{the mouse}, :d \ -k, \text{mouse} \rangle} \text{merge-1} \quad (19-1)$$

$$\frac{\langle \text{eat}, ::=n \ v \ -f, \lambda x. \lambda y. \text{eat}(x)(y) \rangle \quad \langle \text{cheese}, ::=n \ -k, \text{cheese} \rangle}{\langle \text{eat}, :v \ -f, \lambda x. \lambda y. \text{eat}(x)(y) \rangle \langle \text{cheese}, :-k, \text{cheese} \rangle} \text{merge-3} \quad (19-2)$$

$$\frac{\langle \epsilon, ::=v \ +k \ =d \ \text{pred}, \lambda P. \lambda Q. Q(P) \rangle \quad \langle \text{eat}, :v \ -f, \lambda x. \lambda y. \text{eat}(x)(y) \rangle \langle \text{cheese}, :-k, \text{cheese} \rangle}{\langle \epsilon, :+k \ =d \ \text{pred}, \lambda P. \lambda Q. Q(P) \rangle \langle \text{eat}, :-f, \lambda x. \lambda y. \text{eat}(x)(y) \rangle \langle \text{cheese}, :-k, \text{cheese} \rangle} \text{merge-3} \quad (19-3)$$

$$\frac{\langle \epsilon, :+k \ =d \ \text{pred}, \lambda P. \lambda Q. Q(P) \rangle \langle \text{eat}, :-f, \lambda x. \lambda y. \text{eat}(x)(y) \rangle \langle \text{cheese}, :-k, \text{cheese} \rangle}{\langle \text{cheese}, :=d \ \text{pred}, (\lambda P. \lambda Q. Q(P))(\text{cheese}) \rangle \langle \text{eat}, :-f, \lambda x. \lambda y. \text{eat}(x)(y) \rangle} \text{move-1} \quad (19-4)$$

$$\frac{\langle \text{cheese}, :=d \ \text{pred}, (\lambda P. \lambda Q. Q(P))(\text{cheese}) \rangle \langle \text{eat}, :-f, \lambda x. \lambda y. \text{eat}(x)(y) \rangle}{\langle \text{cheese}, :=d \ \text{pred}, \lambda Q. Q(\text{cheese}) \rangle \langle \text{eat}, :-f, \lambda x. \lambda y. \text{eat}(x)(y) \rangle} \lambda\text{-app.} \quad (19-5)$$

$$\frac{\langle \text{cheese}, :=d \ \text{pred}, \lambda Q. Q(\text{cheese}) \rangle \langle \text{eat}, :-f, \lambda x. \lambda y. \text{eat}(x)(y) \rangle \quad \langle \text{the mouse}, :d \ -k, \text{mouse} \rangle}{\langle \text{cheese}, :pred, \lambda Q. Q(\text{cheese}) \rangle \langle \text{eat}, :-f, \lambda x. \lambda y. \text{eat}(x)(y) \rangle \langle \text{the mouse}, :-k, \text{mouse} \rangle} \text{merge-3} \quad (19-6)$$

$$\frac{\langle -s, ::=pred \ +f \ +k \ t, \epsilon \rangle \quad \langle \text{cheese}, :pred, \lambda Q. Q(\text{cheese}) \rangle \langle \text{eat}, :-f, \lambda x. \lambda y. \text{eat}(x)(y) \rangle \langle \text{the mouse}, :-k, \text{mouse} \rangle}{\langle -s \ \text{cheese}, :+f \ +k \ t, \lambda Q. Q(\text{cheese}) \rangle \langle \text{eat}, :-f, \lambda x. \lambda y. \text{eat}(x)(y) \rangle \langle \text{the mouse}, :-k, \text{mouse} \rangle} \text{merge-1} \quad (19-7)$$

$$\frac{\langle -s \ \text{cheese}, :+f \ +k \ t, \lambda Q. Q(\text{cheese}) \rangle \langle \text{eat}, :-f, \lambda x. \lambda y. \text{eat}(x)(y) \rangle \langle \text{the mouse}, :-k, \text{mouse} \rangle}{\langle \text{eat-s cheese}, :+k \ t, (\lambda Q. Q(\text{cheese}))(\lambda x. \lambda y. \text{eat}(x)(y)) \rangle \langle \text{the mouse}, :-k, \text{mouse} \rangle} \text{move-1} \quad (19-8)$$

$$\frac{\langle \text{eats cheese}, :+k \ t, (\lambda Q. Q(\text{cheese}))(\lambda x. \lambda y. \text{eat}(x)(y)) \rangle \langle \text{the mouse}, :-k, \text{mouse} \rangle}{\langle \text{eats cheese}, :+k \ t, (\lambda x. \lambda y. \text{eat}(x)(y))(\text{cheese}) \rangle \langle \text{the mouse}, :-k, \text{mouse} \rangle} \lambda\text{-app.} \quad (19-9)$$

$$\frac{\langle \text{eats cheese}, :+k \ t, (\lambda x. \lambda y. \text{eat}(x)(y))(\text{cheese}) \rangle \langle \text{the mouse}, :-k, \text{mouse} \rangle}{\langle \text{eats cheese}, :+k \ t, \lambda y. \text{eat}(\text{cheese})(y) \rangle \langle \text{the mouse}, :-k, \text{mouse} \rangle} \lambda\text{-app.} \quad (19-10)$$

$$\frac{\langle \text{eats cheese}, :+k \ t, \lambda y. \text{eat}(\text{cheese})(y) \rangle \langle \text{the mouse}, :-k, \text{mouse} \rangle}{\langle \text{the mouse eats cheese}, :t, (\lambda y. \text{eat}(\text{cheese})(y))(\text{mouse}) \rangle} \text{move-1} \quad (19-11)$$

$$\frac{\langle \text{the mouse eats cheese}, :t, (\lambda y. \text{eat}(\text{cheese})(y))(\text{mouse}) \rangle}{\langle \text{the mouse eats cheese}, :t, \text{eat}(\text{cheese})(\text{mouse}) \rangle} \lambda\text{-app.} \quad (19-12)$$

$$\frac{\langle \epsilon, ::=t \ c, \epsilon \rangle \quad \langle \text{the mouse eats cheese}, :t, \text{eat}(\text{cheese})(\text{mouse}) \rangle}{\langle \text{the mouse eats cheese}, :c, \text{eat}(\text{cheese})(\text{mouse}) \rangle} \text{merge-1.} \quad (19-13)$$

We adopt a shallow semantic model, where the universe of discourse only contains two individuals, the mouse and a piece of cheese⁵. Then, the lexicon (**Table 2**) is interpreted as follows. Since all entries are contained in the MG lexicon, they are of category “::.” There are two nouns (n), *mouse* and *cheese* with their respective semantics as individual constants, *mouse* and *cheese*. In contrast to *mouse*, the latter possesses a licensee $-k$ for case marking. The same holds for the determiner *the* selecting a noun ($=n$) as its complement to form a determiner phrase d which also requires case assignment ($-k$) afterwards. The verb (v) *eat* selects a noun as a complement and has to be moved for inflection $-f$. Its compositional

In the first step, (19-1), the determiner *the* takes the noun *mouse* as its complement to form a determiner phrase d that requires licensing through case marking afterwards. In step 2, the finite verb *eat* selects the noun *cheese* as direct object, thus forming a verbal phrase v . As there remain unchecked features, only merge-3 applies yielding a minimalist expression, i.e. a sequence of signs. In step 3, the phonetically empty predicate *pred* merges with the formerly built verbal phrase. Since *pred* assigns accusative case, the direct object is moved in (19-4) toward the first position through case marking by simultaneously concatenating the respective lambda terms. Thus, lambda application entails the expression in step 5. Then, in step 6, the predicate selects its subject, the formerly construed determiner phrase. In the seventh step, (19-7), the present-tense suffix unifies with the predicate, entailing an inflection

⁵Moreover, we abstract our analysis from temporal and numeral semantics and also from the intricacies of the semantics of noun phrases in the present exposition.

TABLE 2 | UMT minimalist lexicon for example grammar (Figure 2).

$\langle \text{mouse}, ::n, \text{mouse} \rangle$	$\langle \text{cheese}, ::n \text{ } -k, \text{cheese} \rangle$
$\langle \text{the}, ::n \text{ } d \text{ } -k, \epsilon \rangle$	$\langle \text{eat}, ::n \text{ } v \text{ } -f, \lambda x. \lambda y. \text{eat}(x)(y) \rangle$
$\langle -s, ::\text{pred } +f \text{ } +k \text{ } t, \epsilon \rangle$	$\langle \epsilon, ::v \text{ } +k \text{ } =d \text{ } \text{pred}, \lambda P. \lambda Q. Q(P) \rangle$
$\langle \epsilon, ::t \text{ } c, \epsilon \rangle$	

phrase *pred*, whose verb is moved into the first position in step 8, thereby yielding the inflected verb *eat-s*. In steps 9 and 10 two lambda applications result into the correct semantics, already displayed in **Figure 2B**. Step 11 assigns nominative case to the subject through movement into specifier position. A further lambda application in step 12 yields the intended interpretation of predicate logics. Finally, in step 13, the syntactic type *t* is converted into *c* to obtain the proper start symbol of the grammar.

Derivations such as (19) are essential for MG. However, their computation is neither incremental nor predictive. Therefore, they are not suitable for natural language processing in their present form of data-driven bottom-up processing. A number of different parsing architectures have been suggested in the literature to remedy this problem (Harkema, 2001; Mainguy, 2010; Stabler, 2011b; Stanojević and Stabler, 2018). From a psycholinguistic point of view, predictive parsing appears most plausible, because a cognitive agent should be able to make informed guesses about a speaker's intents as early as possible, without waiting for the end of an utterance (Hale, 2011). This makes either an hypothesis-driven top-down parser, or a mixed-strategy left-corner parser desirable also for language engineering applications. In this section, we briefly describe a bidirectional utterance-meaning transducer (UMT) for MG that is based upon Stabler (2011b)'s top-down recognizer as outlined earlier by beim Graben et al. (2019a). Its generalization toward the recent left-corner parser (Stanojević and Stabler, 2018) is straightforward.

The central object for MG language processing is the *derivation tree* obtained from a bottom-up derivation as in (19). **Figure 3** depicts this derivation tree, where we present a comma-separated sequence of exponents for the sake of simplicity. Additionally, every node is addressed by an index tuple that is computed according to Stabler (2011b)'s algorithm.

Pursuing the tree paths in **Figure 3** from the bottom to the top, provides exactly the derivation (19). However, reading it from the top toward the bottom allows for an interpretation in terms of *multiple context-free grammars* (Seki et al., 1991; Michaelis, 2001) (MCFG) where categories are *n*-ary predicates over string exponents. Like in context-free grammars, every branching in the derivation tree (**Figure 3**) leads to one phrase structure rule in the MCFG. Thus, the MCFG enumerated in (20) codifies the MG (**Table 2**).

$$\langle :c \rangle(e_0 e_1) \leftarrow \langle ::t \text{ } c \rangle(e_0) \quad \langle :t \rangle(e_1) \quad (20-1)$$

$$\langle :t \rangle(e_1 e_0) \leftarrow \langle :+k \text{ } t, \text{ } -k \rangle(e_0, e_1) \quad (20-2)$$

$$\langle :+k \text{ } t, \text{ } -k \rangle(e_1 e_0, e_2) \leftarrow \langle :+f \text{ } +k \text{ } t, \text{ } -f, \text{ } -k \rangle(e_0, e_1, e_2) \quad (20-3)$$

$$\langle :+f \text{ } +k \text{ } t, \text{ } -f, \text{ } -k \rangle(e_0 e_1, e_2, e_3) \leftarrow \langle ::\text{pred } +f \text{ } +k \text{ } t \rangle(e_0)$$

$$\langle : \text{pred}, \text{ } -f, \text{ } -k \rangle(e_1, e_2, e_3) \quad (20-4)$$

$$\langle : \text{pred}, \text{ } -f, \text{ } -k \rangle(e_0, e_1, e_2) \leftarrow \langle : =d \text{ } \text{pred}, \text{ } -f \rangle(e_0, e_1) \quad \langle : d \text{ } -k \rangle(e_2) \quad (20-5)$$

$$\langle : =d \text{ } \text{pred}, \text{ } -f \rangle(e_2 e_0, e_1) \leftarrow \langle : +k \text{ } =d \text{ } \text{pred}, \text{ } -f, \text{ } -k \rangle(e_0, e_1, e_2) \quad (20-6)$$

$$\langle : +k \text{ } =d \text{ } \text{pred}, \text{ } -f, \text{ } -k \rangle(e_0, e_1, e_2) \leftarrow \langle ::v \text{ } +k \text{ } =d \text{ } \text{pred} \rangle(e_0)$$

$$\langle :v \text{ } -f, \text{ } -k \rangle(e_1, e_2) \quad (20-7)$$

$$\langle :v \text{ } -f, \text{ } -k \rangle(e_0, e_1) \leftarrow \langle ::n \text{ } v \text{ } -f \rangle(e_0) \quad \langle ::n \text{ } -k \rangle(e_1) \quad (20-8)$$

$$\langle : d \text{ } -k \rangle(e_0 e_1) \leftarrow \langle ::n \text{ } d \text{ } -k \rangle(e_0) \quad \langle ::n \rangle(e_1) \quad (20-9)$$

$$\langle ::n \rangle(\text{mouse}) \quad (20-10)$$

$$\langle ::n \text{ } -k \rangle(\text{cheese}) \quad (20-11)$$

$$\langle ::n \text{ } d \text{ } -k \rangle(\text{the}) \quad (20-12)$$

$$\langle ::n \text{ } v \text{ } -f \rangle(\text{eat}) \quad (20-13)$$

$$\langle ::\text{pred } +f \text{ } +k \text{ } t \rangle(-s) \quad (20-14)$$

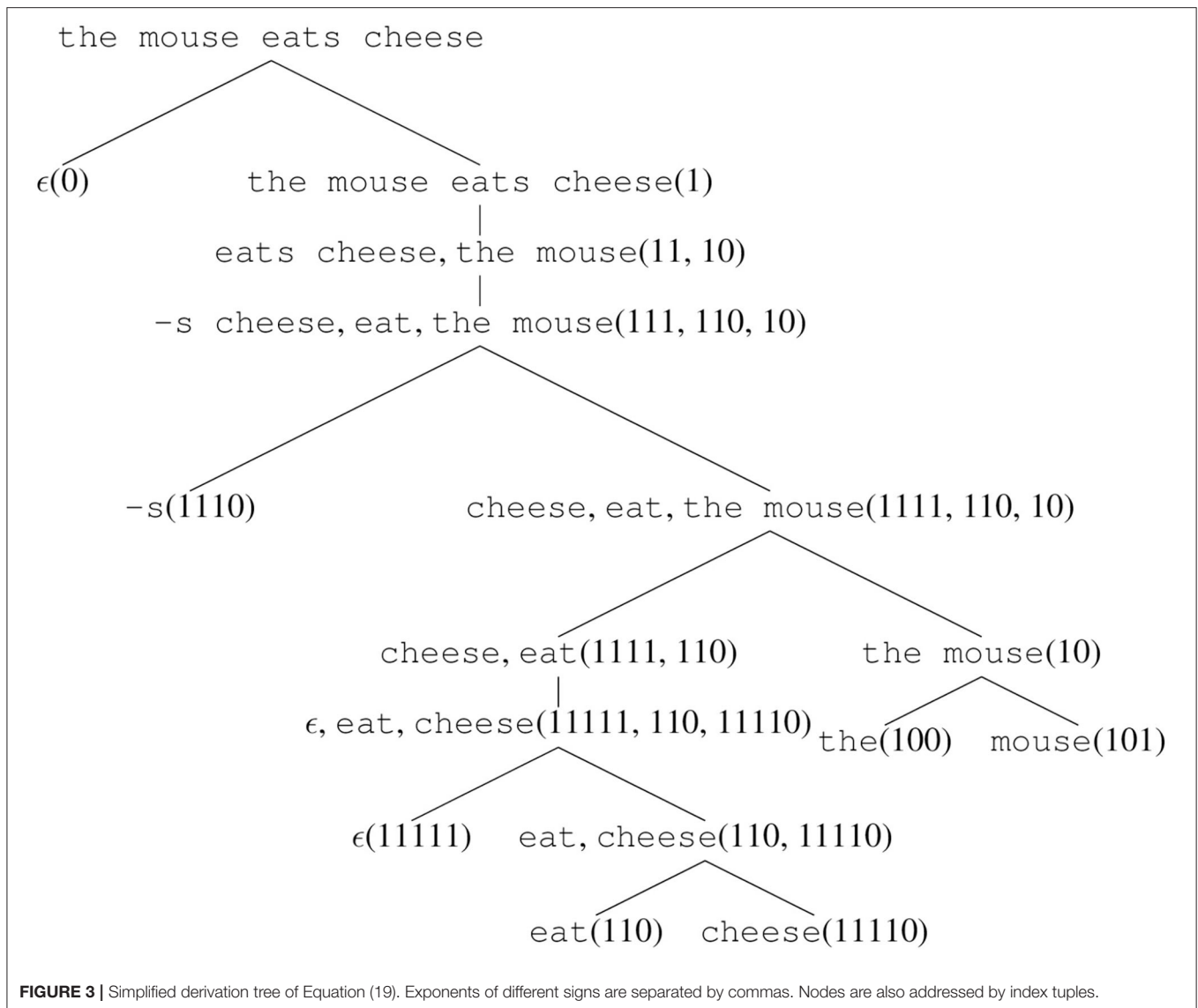
$$\langle ::v \text{ } +k \text{ } =d \text{ } \text{pred} \rangle(\epsilon) \quad (20-15)$$

$$\langle ::t \text{ } c \rangle(\epsilon) \quad (20-16)$$

In (20), the angular brackets enclose the MCFG categories that are obviously formed by tuples of MG categories and syntactic types. These categories have the same number of string arguments e_k as prescribed in the type tuples. Because MCFG serve only for syntactic parsing in our UMT, we deliberately omit the semantic terms here; they are reintroduced below. The MCFG rules (20-1–20-9) are directly obtained from the derivation tree (**Figure 3**) by reverting the merge and move operations of (19) through their “unmerge” and “unmove” counterparts (Harkema, 2001). The MCFG axioms, i. e. the lexical rules (20-10 – 20-16), are reformulations of the entries in the MG lexicon (**Table 2**).

The UMT's language production module finds a semantic representation of an intended utterance in form of a Schönfinkel-Curry (Schönfinkel, 1924; Lohnstein, 2011) formula of predicate logic, such as *eat(cheese)(mouse)*, for instance. According to **Figure 2B** this is a hierarchical data structure that can control the MG derivation (19). Thus, the cognitive agent accesses its mental lexicon, either through **Table 2** or its MCFG twin (20) in order to retrieve the linguistic signs for the denotations *eat*, *cheese*, and *mouse*. Then, the semantic tree (**Figure 2B**) governs the correct derivation (19) up to lexicon entries that are phonetically empty. These must occasionally be queried from the data base whenever required. At the end of the derivation the computed exponent *the mouse eats cheese* is uttered.

The language understanding module of our UMT, by contrast, comprises three memory tapes: the input sequence, a syntactic priority queue, and also a semantic priority queue. Input tape and syntactic priority queue together constitute Stabler (2011b)'s priority queue top-down parser. Yet, in order to compute the meaning of an utterance in the semantic priority queue, we slightly depart from the original proposal by omitting the simplifying trim function. **Table 3** presents the temporal



evolution of the top-down recognizer's configurations while processing the utterance *the mouse eats cheese*.

The parser is initialized with the input string to be processed and the MCFG start symbol $\langle :c \rangle(\epsilon)$ —corresponding to the MG start symbol c —at the top of the priority queue. For each rule of the MCFG (20), the algorithm replaces its string variables by an index tuple that addresses the corresponding nodes in the derivation tree (Figure 3) (Stabler, 2011b). These indices allow for an ordering relation where shorter indices are smaller than longer ones, while indices of equal length are ordered lexicographically. As a consequence, the MCFG axioms in (20) become ordered according to their temporal appearance in the utterance. Using the notation of the derivation tree (Figure 3), we get

$$\begin{aligned} \text{the}(100) &< \text{mouse}(101) < \text{eat}(110) < -s(1110) \\ &< \text{cheese}(11110). \end{aligned}$$

Hence, index sorting ensures incremental parsing.

Besides the occasional sorting of the syntactic priority queue, the automaton behaves as a conventional context-free top-down parser. When the first item in the queue is an MCFG category appearing at the left hand side of an MCFG rule, this item is *expanded* into the right hand side of that rule. When the first item in the queue is a predicted MCFG axiom whose exponent also appears on top of the input tape, this item is *scanned* from the input and thereby removed from queue and input simultaneously. Finally, if queue and input both contain only the empty word, the utterance has been successfully recognized and the parser terminates in the *accepting* state.

Interestingly, the index formalism leads to a straightforward implementation of the UMT's semantic parser as well. The derivation tree (Figure 3) reveals that the index length correlates with the tree depth. In our example, the items $\langle ::n -k \rangle(11110)$ and $\langle ::=v +k =d \text{ pred} \rangle(11111)$ in the priority queue have

TABLE 3 | MG top-down parse of the mouse eats cheese.

Step	Input	Syntactic queue	Operation
1.	the mouse eats cheese	$\langle :c \rangle (\epsilon)$	expand (20-1)
2.	the mouse eats cheese	$\langle ::=t \ c \rangle (0) \langle :t \rangle (1)$	scan (20-16)
3.	the mouse eats cheese	$\langle :t \rangle (1)$	expand (20-2)
4.	the mouse eats cheese	$\langle :+k \ t, \ -k \rangle (11,10)$	expand (20-3)
5.	the mouse eats cheese	$\langle :+f \ +k \ t, \ -f, \ -k \rangle (111,110,10)$	expand (20-4)
6.	the mouse eats cheese	$\langle ::=pred \ +f \ +k \ t \rangle (1110) \langle :pred, \ -f, \ -k \rangle (1111,110,10)$	sort
7.	the mouse eats cheese	$\langle :pred, \ -f, \ -k \rangle (1111,110,10) \langle ::=pred \ +f \ +k \ t \rangle (1110)$	expand (20-5)
8.	the mouse eats cheese	$\langle :d \ pred, \ -f \rangle (1111,110) \langle :d \ -k \rangle (10) \langle ::=pred \ +f \ +k \ t \rangle (1110)$	sort
9.	the mouse eats cheese	$\langle :d \ -k \rangle (10) \langle :d \ pred, \ -f \rangle (1111,110) \langle ::=pred \ +f \ +k \ t \rangle (1110)$	expand (20-9)
10.	the mouse eats cheese	$\langle ::=n \ d \ -k \rangle (100) \langle ::n \rangle (101) \langle :d \ pred, \ -f \rangle (1111,110) \langle ::=pred \ +f \ +k \ t \rangle (1110)$	scan (20-12)
11.	mouse eats cheese	$\langle ::n \rangle (101) \langle :d \ pred, \ -f \rangle (1111,110) \langle ::=pred \ +f \ +k \ t \rangle (1110)$	scan (20-10)
12.	eats cheese	$\langle :d \ pred, \ -f \rangle (1111,110) \langle ::=pred \ +f \ +k \ t \rangle (1110)$	expand (20-6)
13.	eats cheese	$\langle :+k \ d \ pred, \ -f, \ -k \rangle (11111,110,11110) \langle ::=pred \ +f \ +k \ t \rangle (1110)$	expand (20-7)
14.	eats cheese	$\langle ::=v \ +k \ d \ pred \rangle (11111) \langle :v \ -f, \ -k \rangle (110,11110) \langle ::=pred \ +f \ +k \ t \rangle (1110)$	sort
15.	eats cheese	$\langle :v \ -f, \ -k \rangle (110,11110) \langle ::=pred \ +f \ +k \ t \rangle (1110) \langle ::=v \ +k \ d \ pred \rangle (11111)$	expand (20-8)
16.	eats cheese	$\langle ::=n \ v \ -f \rangle (110) \langle ::n \ -k \rangle (11110) \langle ::=pred \ +f \ +k \ t \rangle (1110) \langle ::=v \ +k \ d \ pred \rangle (11111)$	sort
17.	eats cheese	$\langle ::=n \ v \ -f \rangle (110) \langle ::=pred \ +f \ +k \ t \rangle (1110) \langle ::n \ -k \rangle (11110) \langle ::=v \ +k \ d \ pred \rangle (11111)$	scan (20-13)
18.	-s cheese	$\langle ::=pred \ +f \ +k \ t \rangle (1110) \langle ::n \ -k \rangle (11110) \langle ::=v \ +k \ d \ pred \rangle (11111)$	scan (20-14)
19.	cheese	$\langle ::n \ -k \rangle (11110) \langle ::=v \ +k \ d \ pred \rangle (11111)$	scan (20-11)
20.	ϵ	$\langle ::=v \ +k \ d \ pred \rangle (11111)$	scan (20-15)
21.	ϵ	ϵ	accept

TABLE 4 | Semantic processing of the mouse eats cheese.

Step	Input	Semantic queue	Operation
1.	the mouse eats cheese	ϵ	scan (20-16)
2.	the mouse eats cheese	$\langle \epsilon \rangle (0)$	apply
3.	the mouse eats cheese	ϵ	scan (20-12)
4.	mouse eats cheese	$\langle \epsilon \rangle (100)$	apply
5.	mouse eats cheese	ϵ	scan (20-10)
6.	eats cheese	$\langle \text{mouse} \rangle (101)$	scan (20-13)
7.	-s cheese	$\langle \text{mouse} \rangle (101) \langle \lambda x. \lambda y. \text{eat}(x)(y) \rangle (110)$	scan (20-14)
8.	cheese	$\langle \text{mouse} \rangle (101) \langle \lambda x. \lambda y. \text{eat}(x)(y) \rangle (110) \langle \epsilon \rangle (1110)$	apply
9.	cheese	$\langle \text{mouse} \rangle (101) \langle \lambda x. \lambda y. \text{eat}(x)(y) \rangle (110)$	scan (20-11)
10.	ϵ	$\langle \text{mouse} \rangle (101) \langle \lambda x. \lambda y. \text{eat}(x)(y) \rangle (110) \langle \text{cheese} \rangle (11110)$	scan (20-15)
11.	ϵ	$\langle \text{mouse} \rangle (101) \langle \lambda x. \lambda y. \text{eat}(x)(y) \rangle (110) \langle \text{cheese} \rangle (11110) \langle \lambda P. \lambda Q. Q(P) \rangle (11111)$	sort
11.	ϵ	$\langle \lambda P. \lambda Q. Q(P) \rangle (11111) \langle \text{cheese} \rangle (11110) \langle \lambda x. \lambda y. \text{eat}(x)(y) \rangle (110) \langle \text{mouse} \rangle (101)$	apply
12.	ϵ	$\langle \lambda Q. Q(\text{cheese}) \rangle (1111) \langle \lambda x. \lambda y. \text{eat}(x)(y) \rangle (110) \langle \text{mouse} \rangle (101)$	apply
13.	ϵ	$\langle \lambda x. \lambda y. \text{eat}(x)(y) \rangle (\text{cheese}) (111) \langle \text{mouse} \rangle (101)$	apply
14.	ϵ	$\langle \lambda y. \text{eat}(\text{cheese})(y) \rangle (111) \langle \text{mouse} \rangle (101)$	apply
15.	ϵ	$\langle \text{eat}(\text{cheese})(\text{mouse}) \rangle (11)$	understand

the longest indices. These correspond precisely to the lambda terms *cheese* and $\lambda P. \lambda Q. Q(P)$, respectively, that are unified by lambda application in derivation step (19-5). Moreover, also the semantic analysis in **Figure 2B** illustrates that the deepest nodes are semantically unified first.

Every time, when the syntactic parser scans a correctly predicted item from the input tape, this item is removed from both input tape and syntactic priority queue. Simultaneously, the semantic content of its sign is pushed on top of the semantic

priority queue, yet preserving its index. When some or all semantic items are stored in the queue, they are sorted in *reversed* index order to get highest semantic priority on top of the queue.

Table 4 illustrates the semantic parsing for the given example.

In analogy to the syntactic recognizer, the semantic parser operates in similar modes. Both processors share their common *scan* operation. In contrast to the syntactic parser which sorts indices in ascending order, the semantic module *sorts* them in descending order for operating on the deepest nodes in

TABLE 5 | Minimalist lexicon–communication.

$\langle \text{contains}, ::n =n v, \lambda q. \lambda p. \text{contain}(q)(p) \rangle$	$\langle \text{cheese}, ::n, \text{cheese} \rangle$
$\langle \text{lies}, ::p =n v, \lambda q. \lambda p. \text{lie}(q)(p) \rangle$	$\langle \text{field}, ::n, \text{field} \rangle$
$\langle (x, y), ::p -f, (x, y) \rangle$	$\langle \epsilon, ::n =p m -g, \lambda P. \lambda p. P(p) \rangle$
$\langle \epsilon, ::m +f +g n, \lambda P. \lambda Q. QP \rangle$	$\langle \text{in}, ::n p, \epsilon \rangle$

the derivation tree first. Most of the time, it attempts lambda application (*apply*) which is always preferred for ϵ -items on the queue. When apply has been sufficiently performed upon a term, the last index number is removed from its index (sometimes it might also be necessary to exchange two items for lambda application). Finally, the semantic parser terminates in the *understanding* state.

5.2. Interpretation

After preparing the UMT, we come back to our particular mouse-maze example. As previously explained, the agent could find its target objects by physically exploring its state space through interaction. Yet, another possibility is linguistic communication where an operator may inform the agent about the correct position of a target. For this case, we present another MG in **Table 5**. It contains the linguistic signs that can be used to compose the meaning of a message.

Our “communication” lexicon comprises two verbs, *contains* and *lies*, of the basic type v , which we want to consider here as the start symbol of the MG. The verb *contains* is transitive and therefore corresponds to a

binary predicate, as indicated above by the lambda term in the Schönfinkel representation (Lohnstein, 2011). Syntactically, this binary predicate is expressed by two selectors $=n$, whereas in linguistics we have to take into account, that a transitive verb first selects a direct object and subsequently its subject (both nouns of the type n “noun”). The verb *lies* is intransitive, but can be modified by a prepositional phrase through a selector ($=p$). There are two nouns in the lexicon: *field* and *cheese*, their semantics are the corresponding individual names. The *field* coordinates (x, y) in the labyrinth are expressed by the exponent (x, y) , whose syntactic function is an adjunct of the basic type p (prepositional phrase). This must be moved once in the course of the structure development, for which the licensee $-f$ is intended. The adjunction itself is triggered by a phonetically empty entry of the basic type m (modifier). This has two selectors: $=n$ selects the noun phrase to be modified, while $=p$ selects the adjunct prepositional phrase. Then $-g$ licenses a movement. The semantics of the adjunction here expressed in a simplified manner as the predication $\lambda P. \lambda p. P(p)$. That is, there is an object p which has the property P . Another phonetically empty expression selects a modified noun phrase ($=m$), then licenses two movements ($+f +g$) and generates itself a noun phrase of the base type n . Semantically, an argument movement is canceled again by the exchange operator $\lambda P. \lambda Q. QP$. Finally, there is a preposition *in* of the basic type p , which also selects a noun phrase for adjunction ($=n$). Instead of formalizing their locative semantics in concrete terms, we simplify them with an empty lambda expression ϵ . In order to reduce the effort for the exploration phase, we can send the following message “*field (x, y) contains cheese*” to the mouse. This utterance can be processed by our UMT above, for which we present the minimalist bottom-up derivation as follows.

$$\begin{array}{c}
 \frac{\langle \text{contains}, ::n =n v, \lambda q. \lambda p. \text{contain}(q)(p) \rangle \quad \langle \text{cheese}, ::n, \text{cheese} \rangle}{\langle \text{contains cheese}, ::n v, (\lambda q. \lambda p. \text{contain}(q)(p))(\text{cheese}) \rangle} \text{merge-1} \\
 \\
 \frac{\langle \epsilon, ::n =p m -g, \lambda P. \lambda p. P(p) \rangle \quad \langle \text{field}, ::n, \text{field} \rangle}{\langle \text{field}, ::p m -g, (\lambda P. \lambda p. P(p))(\text{field}) \rangle} \text{merge-1} \\
 \\
 \frac{\langle \text{field}, ::p m -g, \lambda p. \text{field}(p) \rangle \quad \langle (x, y), :p -f, (x, y) \rangle}{\langle \text{field}, :m -g, \lambda p. \text{field}(p) \rangle \langle (x, y), :-f, (x, y) \rangle} \text{merge-3} \\
 \\
 \frac{\langle \epsilon, ::m +f +g n, \lambda P. \lambda Q. QP \rangle \quad \langle \text{field}, :m -g, \lambda p. \text{field}(p) \rangle \langle (x, y), :-f, (x, y) \rangle}{\langle \epsilon, :+f +g n, \lambda P. \lambda Q. QP \rangle \langle \text{field}, :-g, \lambda p. \text{field}(p) \rangle \langle (x, y), :-f, (x, y) \rangle} \text{merge-3} \\
 \\
 \frac{\langle \epsilon, :+f +g n, \lambda P. \lambda Q. QP \rangle \langle \text{field}, :-g, \lambda p. \text{field}(p) \rangle \langle (x, y), :-f, (x, y) \rangle}{\langle (x, y), :+g n, (\lambda P. \lambda Q. QP)((x, y)) \rangle \langle \text{field}, :-g, \lambda p. \text{field}(p) \rangle} \text{move-1} \\
 \\
 \frac{\langle (x, y), :+g n, \lambda Q. Q((x, y)) \rangle \langle \text{field}, :-g, \lambda p. \text{field}(p) \rangle}{\langle \text{field } (x, y), :n, (\lambda Q. Q((x, y))) (\lambda p. \text{field}(p)) \rangle} \text{move-1} \\
 \\
 \frac{\langle \text{contains cheese}, ::n v, \lambda p. \text{contain}(\text{cheese})(p) \rangle \quad \langle \text{field } (x, y), :n, \text{field}((x, y)) \rangle}{\langle \text{field } (x, y) \text{ contains cheese}, :v, (\lambda p. \text{contain}(\text{cheese})(p))(\text{field}((x, y))) \rangle} \text{merge-2}
 \end{array}$$

Thereby, the utterance “field (x,y) contains cheese” is recognized in the exponent, so that its semantics can be interpreted in terms of model theory as a predicate logic formula after a last lambda application $(\lambda p.\text{contain}(\text{cheese})(p))(\text{field}((x,y))) = \text{contain}(\text{cheese})(\text{field}((x,y)))$. The denotation of this predicate assignment results to

$$[\text{contain}(\text{cheese})(\text{field}((x,y)))] = ([\text{cheese}], [\text{field}((x,y))]) \in [\text{contain}], \quad (21)$$

This corresponds to a representational data structure which can be inserted in an associated relational scheme.

5.3. Articulation

For speech production, we start from the fact “field (x,y) contains cheese” and show the possibility of stylistic creativity. First, we note that the denotation of *contain* is coextensive with that of *lie*, i.e. $[\text{contain}] = [\text{lie}]$ applies. Based on this capability an agent is able to express what it has understood in its own words. In this way, the adjustment of common ideas or concepts required for successful communication can take place. If the agent wants to articulate such an utterance, it must first translate the facts into a predicate logic formula and has to consider, that the order of arguments of the subject and the direct object must be swapped: $\text{lie}(\text{field}((x,y)))(\text{cheese})$. A query in the minimalist lexicon then leads to the following derivation, whereby we reuse the previously derived expression $(\text{field} (x,y), :n, \text{field}((x,y)))$ in the sense of dynamic programming.

$$\begin{array}{c} \frac{\langle \text{in}, ::n \text{ p}, \epsilon \rangle \quad \langle \text{field} (x,y), :n, \text{field}((x,y)) \rangle}{\langle \text{in field} (x,y), :p, \text{field}((x,y)) \rangle} \text{merge-1} \\ \frac{\langle \text{lies}, ::p =n \text{ v}, \lambda q. \lambda p. \text{lie}(q)(p) \rangle \quad \langle \text{in field} (x,y), :p, \text{field}((x,y)) \rangle}{\langle \text{lies in field} (x,y), :n \text{ v}, (\lambda q. \lambda p. \text{lie}(q)(p))(\text{field}((x,y))) \rangle} \text{merge-1} \\ \frac{\langle \text{lies in field} (x,y), :n \text{ v}, \lambda p. \text{lie}(\text{field}((x,y)))(p) \rangle \quad \langle \text{cheese}, ::n, \text{cheese} \rangle}{\langle \text{cheese lies in field} (x,y), :v, (\lambda p. \text{lie}(\text{field}((x,y)))(p))(\text{cheese}) \rangle} \text{merge-2} \end{array}$$

The present derivation finally comprises after a final lambda application the intended semantics $(\lambda p. \text{lie}(\text{field}((x,y)))(p))(\text{cheese}) = \text{lie}(\text{field}((x,y)))(\text{cheese})$ and leads also to the synonymous utterance “cheese lies in field (x,y).”

5.4. Knowledge Integration

In the previous sections we described how the information of symbolic sequences can be transformed into logically processable knowledge and vice versa by interaction or communication, respectively. In order to process knowledge, it has to be saved and retrieved again. These operations are essential characteristics of a universal algorithmic system with which calculation rules can be implemented. Based on such a knowledge processing system a cognitive agent is able to pursue goals, understand situations, select cost optimal actions and to predict its consequences. Further, the agent should be able to learn behavioral relations and to adapt to changing environmental conditions. These requirements can be satisfied more efficiently if the required

knowledge is structured and available as a (dynamic) model of the external world. To this end, we apply the state space model, with which we can describe the relations $R_{\text{Situation}} \subseteq \text{States} \times \mathcal{O}$ and $R_{\text{Movement}} \subseteq \text{States} \times \mathcal{A} \times \text{States}$, where the set *States* corresponds to the set \mathcal{Z} of the state space model. Based on these relations predictions about the next state and the associated observations can be made. Prediction errors can either be used to recognize changes in the environmental conditions or in the realization of the selected actions, which can initiate an adaptation or diagnosis process. The states $z \in \mathcal{Z}$ correspond to the maze fields and have a finer inner structure based on ordered pairs $z = (x, y) \in X \times Y$, which are associated with the two-dimensional coordinates of the maze. They can be measured by sensors. The same applies to the elements $a \in \mathcal{A}$, where the finer structure is given by $a = (\Delta x, \Delta y) \in \Delta X \times \Delta Y$. The elements $o \in \mathcal{O}$ correspond to the target object types, which can be identified by an object classifier.

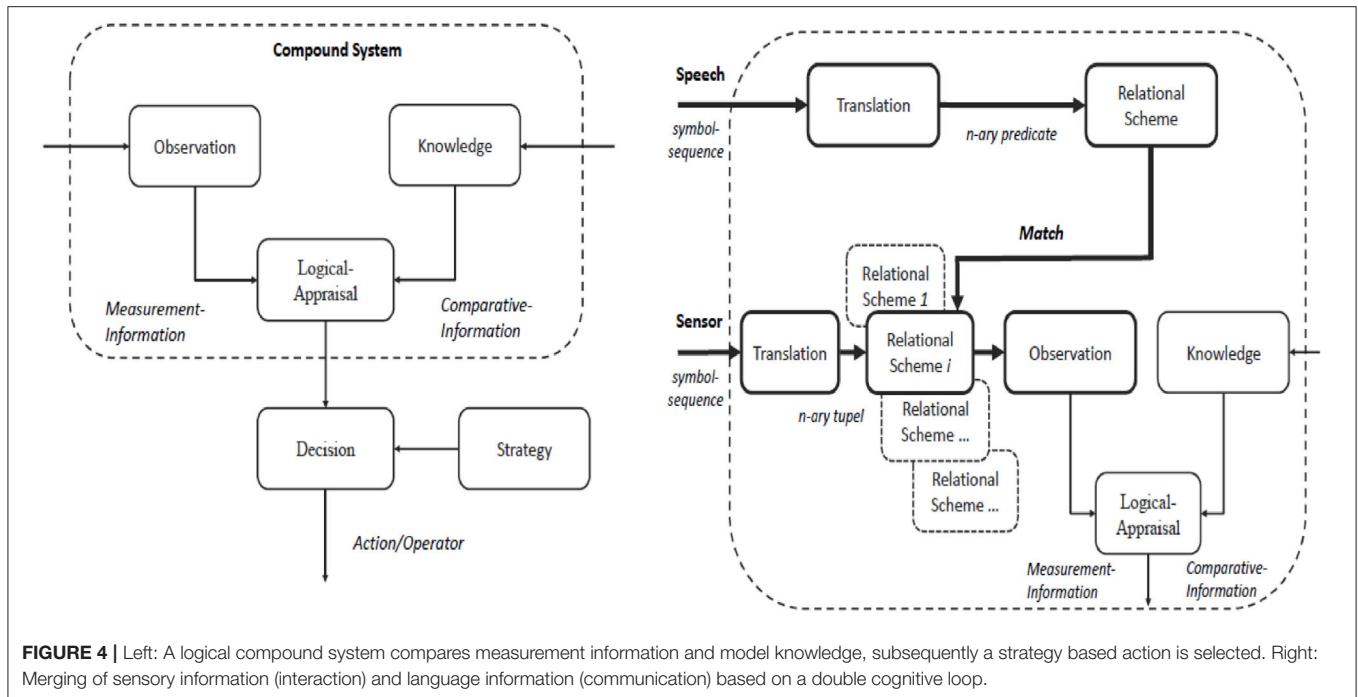
In order to build a dynamic model, the agent must have the ability to analyze experimentally. In the course of this experimentation process (exploration), knowledge about “situations” and “movements” is stored in relational schemes as long as no contradictions occur. Such an exploration process results—mathematically speaking—in a relation $R_{\mathcal{X}} \in \text{Rel}(\mathcal{X})$, whereby the identified relation $R_{\mathcal{X}}$ can be expressed by its characteristic function (e.g., $f_{R_{\text{Situation}}} : X \times Y \times \mathcal{O} \rightarrow \{0, 1\}$), which assigns exactly one truth value to each tuple $\tau \in \text{Typ}(\mathcal{X})$. After completion of the exploration phase, the evaluation of the characteristic functions can be used to answer questions like “What is the case?” or “What does this observation mean?” in a truth-functional sense and therefore leads to a veridical model. Such a model enables the agent to gain knowledge through simulation or inference, which in turn can then be transferred to

reality. To this end, the agent has to compare the representational data structure of the perception (e.g. $\tau_{\text{obs}} = \{X \rightarrow x \in \text{dom}(X), Y \rightarrow y \in \text{dom}(Y), \rightarrow o \in \text{dom}(\mathcal{O})\}$) with some of the learned representational data structures of the associated relational scheme:

$$f_{R_{\mathcal{X}}}(\tau) = \begin{cases} \text{True}, & \text{if } \tau \in R_{\mathcal{X}}, \\ \text{False}, & \text{if } \tau \notin R_{\mathcal{X}}. \end{cases}$$

If matches are found, the agent understands the incoming information. In semiotics, this process is called denotation and can be realized by a logical compound system (see **Figure 4**). If no matches are found, an adaptation or coping process is triggered (Wolff et al., 2015).

In order to apply the dynamic model to the mouse-maze-application properly, we have to specify all attributes and relation schemes of states, situations, actions and movements as well as the corresponding relations that are required: $\text{States} =$



$\{X, Y\} \subset \mathcal{U}$, $R(\text{States}) = (X : \text{dom}(X), Y : \text{dom}(Y))$, $R_{\text{States}} \subseteq \text{Tup}(\text{States})$, $\text{Situation} = \{\text{States}, \mathcal{O}\} \subset \mathcal{U}$, $R(\text{Situation}) = (X : \text{dom}(X), Y : \text{dom}(Y), \mathcal{O} : \text{dom}(\mathcal{O}))$ and $R_{\text{Situation}} \subseteq \text{Tup}(\text{Situation})$. $\text{Action} = \{\Delta X, \Delta Y\} \subset \mathcal{U}$, $R(\text{Action}) = (\Delta X : \text{dom}(\Delta X), \Delta Y : \text{dom}(\Delta Y))$ and $R_{\text{Action}} \subseteq \text{Tup}(\text{Action})$. $\text{Movement} = \{\text{States}, \text{Action}\} \subset \mathcal{U}$, $R(\text{Movement}) = (\text{States} : \text{dom}(\text{States}), \text{Action} : \text{dom}(\text{Action}), \text{States}' : \text{dom}(\text{States}'))$ and $R_{\text{Movement}} \subseteq \text{Tup}(\text{Movement})$.

With these definitions suitable representational data structures are provided that are needed for the transformation and integration processes. However, despite the uniform description of the two cognitive loops using linguistic means, there is an essential difference between interaction and communication. Verbal communication operates primarily with object types and refers to predicate symbols as well. In contrast, the perceptive part of interaction is based on sensors and classifiers only, so that no relations between any object types can be directly encoded (Hausser, 2014). Hence, such relations have to be learned by the agent (Wolff et al., 2018). Yet, in a shared environment these different kinds of information can be used to solve the context problem. For this end, we consider the following example, in which the linguistic message “field(x,y) contains cheese” is translated into the binary predicate $\text{contain}(\text{field}(x,y), \text{cheese})$. In order to obtain the truth value of this statement, it must be checked whether an instance of the object type “cheese” is located at the specified position $z = (x,y)$. That means the agent has to move to this position and take measurements. Then, the corresponding symbol sequence must be transformed into the tuple $\tau_{\text{obs}} = \{X \rightarrow x \in \text{dom}(X), Y \rightarrow y \in \text{dom}(Y), \mathcal{O} \rightarrow o \in \text{dom}(\mathcal{O})\} \in R_{\text{Situation}}$, which can now be compared with the entries of the relation $R_{\text{contain}} \subseteq \times X \times Y \times \mathcal{O}$. Please note

that this relation is associated to the translated predicate and that an interpretation of the sensor values appropriate to the context is only possible if the proper scheme has been selected (Figure 4). Since both relations are designated by different names, it must be ensured that their denotation is identical, i.e. $[\text{contain}] = [\text{Situation}]$ must apply.

6. LANGUAGE ACQUISITION

So far we discussed how a cognitive agent, being either human or an intelligent machine, could produce and understand utterances that are described in terms of minimalist grammar. An MG is given by a mental lexicon as in example (Table 2), encoding a large amount of linguistic expert knowledge. Therefore, it seems unlikely that speech-controlled user interfaces could be build and sold by engineering companies for little expenses.

Yet, it has been shown that MG are effectively learnable in the sense of Gold’s formal learning theory (Gold, 1967). The studies of Bonato and Retoré (2001), Kobele et al. (2002), and Stabler et al. (2003) demonstrated how MG can be acquired by positive examples from linguistic dependence graphs (Nivre, 2003; Boston et al., 2010). The required dependency structures can be extracted from linguistic corpora by means of big data machine learning techniques, such as the expectation maximization (EM) algorithm (Klein and Manning, 2004).

In our terminology, such statistical learning methods only reveal correlations at the exponent level of linguistic signs. By contrast, in the present study we propose an alternative training algorithm that simultaneously analyzes similarities between exponents and semantic terms. Moreover, we exploit both positive and negative examples to obtain a better performance

TABLE 6 | Learned minimalist lexicon X_1 at time $t = 1$.

(the mouse eats cheese, :c, eat(cheese)(mouse))

through reinforcement learning (Skinner, 2015; Sutton and Barto, 2018).

The language learner is a cognitive agent \mathcal{L} in a state X_t , to be identified with \mathcal{L} 's mental lexicon at training time t . At time $t = 0$, X_0 is initialized as a *tabula rasa* with the empty lexicon

$$X_0 \leftarrow \emptyset \quad (22)$$

and exposed to UMPs produced by a teacher \mathcal{T} . Note that we assume \mathcal{T} presenting already complete UMPs and not singular utterances to \mathcal{L} . Thus, we circumvent the *symbol grounding problem* of firstly assigning meanings σ to uttered exponents e (Harnad, 1990), which will be addressed in future research. Moreover, we assume that \mathcal{L} is instructed to reproduce \mathcal{T} 's utterances based on its own semantic understanding. This provides a feedback loop and therefore applicability of reinforcement learning (Skinner, 2015; Sutton and Barto, 2018). For our introductory example, we adopt the simple semantic model from Sect. 5. In each iteration, the teacher utters an UMP that should be learned by the learner.

6.1. First Iteration

Let the teacher \mathcal{T} make the first utterance (2)

$$u_1 = \langle \text{the mouse eats cheese, eat(cheese)(mouse)} \rangle.$$

As long as \mathcal{L} is not able to detect patterns or common similarities in \mathcal{T} 's UMPs, it simply adds new entries directly to its mental lexicon, assuming that all utterances are complex ":" and possessing base type c , i. e. the MG start symbol. Hence, \mathcal{L} 's state X_t evolves according to the update rule

$$X_t \leftarrow X_{t-1} \cup \{ \langle e_t, :c, \sigma_t \rangle \}, \quad (23)$$

when $u_t = \langle e_t, \sigma_t \rangle$ is the UMP presented at time t by \mathcal{T} .

In this way, the mental lexicon X_1 shown in **Table 6** has been acquired at time $t = 1$.

6.2. Second Iteration

Next, let the teacher be uttering another proposition

$$u_2 = \langle \text{the rat eats cheese, eat(cheese)(rat)} \rangle. \quad (24)$$

Looking at u_1, u_2 together, the agent's pattern matching module (cf. van Zaenen, 2001) is able to find similarities between exponents and semantics, underlined in Equation (25).

$$u_1 = \langle \text{the mouse } \underline{\text{eats cheese}}, \underline{\text{eat(cheese)}}(\text{mouse}) \rangle \quad (25-1)$$

$$u_2 = \langle \text{the rat } \underline{\text{eats cheese}}, \underline{\text{eat(cheese)}}(\text{rat}) \rangle. \quad (25-2)$$

TABLE 7 | Learned minimalist lexicon X_2 at time $t = 2$.

(<u>the</u> mouse, :d, mouse)	(<u>the</u> rat, :d, rat)
(eats cheese, :=d c, λy . eat(cheese)(y))	

TABLE 8 | Revised minimalist lexicon X_{21} .

(the, :=n d, ϵ)	(mouse, :=n, mouse)
(rat, :=n, rat)	(eats cheese, :=d c, λy . eat(cheese)(y))

TABLE 9 | Learned minimalist lexicon X_3 at time $t = 3$.

(the, :=n d, ϵ)	(mouse, :=n, mouse)
(rat, :=n, rat)	(cheese, :=n, cheese)
(carrot, :=n, carrot)	(eats, :=n =d c, λx . λy . eat(x)(y))

Thus, \mathcal{L} creates two distinct items for the mouse and the rat, respectively, and carries out lambda abstraction to obtain the updated lexicon X_2 in **Table 7**.

Note that the induced variable symbol y and syntactic types d, c are completely arbitrary and do not have any particular meaning to the agent.

As indicated by underlines in **Table 7**, the exponents the mouse and the rat, could be further segmented through pattern matching, that is not reflected by their semantic counterparts, though. Therefore, a revised lexicon X_{21} , displayed in **Table 8** can be devised.

For closing the reinforcement cycle, \mathcal{L} is supposed to produce utterances upon its own understanding. Thus, we assume that \mathcal{L} wants to express the proposition eat(cheese)(rat). According to our discussion in Sect. 5, the corresponding signs are retrieved from the lexicon X_{21} and processed through a valid derivation leading to the correct utterance the rat eats cheese, that is subsequently endorsed by \mathcal{T} .

6.3. Third Iteration

In the third training session, the teacher's utterance might be

$$u_3 = \langle \text{the mouse eats carrot, eat(carrot)(mouse)} \rangle. \quad (26)$$

Now we have to compare u_3 with the lexicon entry for eats cheese in (27).

$$\langle \text{the mouse } \underline{\text{eats}} \text{ carrot, } \underline{\text{eat}}(\text{carrot})(\text{mouse}) \rangle \quad (27-1)$$

$$\langle \underline{\text{eats}} \text{ cheese, :c, } \lambda y. \underline{\text{eat}}(\text{cheese})(y) \rangle. \quad (27-2)$$

Another lambda abstraction entails the lexicon X_3 in **Table 9**.

Here, the learner assumes that eats is a simple, lexical category without having further evidence as in Sect. 5.

Since \mathcal{L} is instructed to produce well-formed utterances, it could now generate a novel semantic representation, such as eat(carrot)(rat). This leads through data base query from the mental lexicon X_3 to the correct derivation (28) that is rewarded by \mathcal{T} .

TABLE 10 | Learned minimalist lexicon X_4 at time $t = 4$.

$\langle \text{the}, ::n \text{ d}, \epsilon \rangle$	$\langle \text{mouse}, ::n, \text{mouse} \rangle$
$\langle \text{rat}, ::n, \text{rat} \rangle$	$\langle \text{rats}, ::n, \text{rats} \rangle$
$\langle \text{cheese}, ::n, \text{cheese} \rangle$	$\langle \text{carrot}, ::n, \text{carrot} \rangle$
$\langle \text{eats}, ::n =d \text{ c}, \lambda x. \lambda y. \text{eat}(x)(y) \rangle$	$\langle \text{eat}, ::n =d \text{ c}, \lambda x. \lambda y. \text{eat}(x)(y) \rangle$

$$\frac{\langle \text{the}, ::n \text{ d}, \epsilon \rangle \quad \langle \text{rat}, ::n, \text{rat} \rangle}{\langle \text{the rat}, :d, \text{rat} \rangle} \text{merge-1} \quad (28-1)$$

$$\frac{\langle \text{eats}, ::n =d \text{ c}, \lambda x. \lambda y. \text{eat}(x)(y) \rangle \quad \langle \text{carrot}, ::n, \text{carrot} \rangle}{\langle \text{eats carrot}, :d \text{ c}, (\lambda x. \lambda y. \text{eat}(x)(y))(\text{carrot}) \rangle} \text{merge-1} \quad (28-2)$$

$$\frac{\langle \text{eats carrot}, :d \text{ c}, (\lambda x. \lambda y. \text{eat}(x)(y))(\text{carrot}) \rangle}{\langle \text{eats carrot}, :d \text{ c}, \lambda y. \text{eat}(\text{carrot})(y) \rangle} \lambda\text{-appl.} \quad (28-3)$$

$$\frac{\langle \text{eats carrot}, :d \text{ c}, \lambda y. \text{eat}(\text{carrot})(y) \rangle \quad \langle \text{the rat}, :d, \text{rat} \rangle}{\langle \text{the rat eats carrot}, :c, (\lambda y. \text{eat}(\text{carrot})(y))(\text{rat}) \rangle} \text{merge-2} \quad (28-4)$$

$$\frac{\langle \text{the rat eats carrot}, :c, (\lambda y. \text{eat}(\text{carrot})(y))(\text{rat}) \rangle}{\langle \text{the rat eats carrot}, :c, \text{eat}(\text{carrot})(\text{rat}) \rangle} \lambda\text{-appl.} \quad (28-5)$$

6.4. Fourth Iteration

In the fourth iteration, we suppose that \mathcal{T} utters

$$u_4 = \langle \text{the rats eat cheese}, \text{eat}(\text{cheese})(\text{rats}) \rangle \quad (29)$$

that is unified with the previous lexicon X_3 through our pattern matching algorithm to yield X_4 in **Table 10** in the first place.

Underlined are again common strings in exponents or semantics that could entail further revisions of the MG lexicon.

Next, let us assume that \mathcal{L} would express the meaning $\text{eat}(\text{carrot})(\text{rats})$. It could then attempt the following derivation (30).

$$\frac{\langle \text{the}, ::n \text{ d}, \epsilon \rangle \quad \langle \text{rats}, ::n, \text{rats} \rangle}{\langle \text{the rats}, :d, \text{rats} \rangle} \text{merge-1} \quad (30-1)$$

$$\frac{\langle \text{eats}, ::n =d \text{ c}, \lambda x. \lambda y. \text{eat}(x)(y) \rangle \quad \langle \text{carrot}, ::n, \text{carrot} \rangle}{\langle \text{eats carrot}, :d \text{ c}, (\lambda x. \lambda y. \text{eat}(x)(y))(\text{carrot}) \rangle} \text{merge-1} \quad (30-2)$$

$$\frac{\langle \text{eats carrot}, :d \text{ c}, (\lambda x. \lambda y. \text{eat}(x)(y))(\text{carrot}) \rangle}{\langle \text{eats carrot}, :d \text{ c}, \lambda y. \text{eat}(\text{carrot})(y) \rangle} \lambda\text{-appl.} \quad (30-3)$$

$$\frac{\langle \text{eats carrot}, :d \text{ c}, \lambda y. \text{eat}(\text{carrot})(y) \rangle \quad \langle \text{the rats}, :d, \text{rats} \rangle}{\langle \text{the rats eats carrot}, :c, (\lambda y. \text{eat}(\text{carrot})(y))(\text{rats}) \rangle} \text{merge-2} \quad (30-4)$$

$$\frac{\langle \text{the rats eats carrot}, :c, (\lambda y. \text{eat}(\text{carrot})(y))(\text{rats}) \rangle}{\langle \text{the rats eats carrot}, :c, \text{eat}(\text{carrot})(\text{rats}) \rangle} \lambda\text{-appl.} \quad (30-5)$$

However, uttering the rats eats carrot will probably be rejected by the teacher \mathcal{T} because of the grammatical number agreement error, thus causing punishment by \mathcal{T} . As a consequence, \mathcal{L} has to find a suitable revision of its lexicon X_4 that is guided by the underlined matches in **Table 10**.

To this aim, the agent first modifies X_4 as given in **Table 11**.

In **Table 11** the entries for mouse and rat have been updated by a number licensee $-a$ (for *Anzahl*). Moreover, the entry for the now selects a number type $=\text{num}$ instead of a noun. Even more crucially, two novel entries of number type num have been added: a phonetically empty item $\langle \epsilon, ::n +a \text{ num}, \epsilon \rangle$ selecting a noun $=n$ and licensing number movement $+a$, and an item for the plural suffix $\langle -s, ::n +a \text{ num}, \epsilon \rangle$ with the same feature sequence.

TABLE 11 | Revised minimalist lexicon X_{41} .

$\langle \text{the}, ::\text{num } d, \epsilon \rangle$	$\langle \text{mouse}, ::n -a, \text{mouse} \rangle$
$\langle \text{rat}, ::n -a, \text{rat} \rangle$	$\langle \epsilon, ::n +a \text{ num}, \epsilon \rangle$
$\langle -s, ::n +a \text{ num}, \epsilon \rangle$	$\langle \text{cheese}, ::n, \text{cheese} \rangle$
$\langle \text{carrot}, ::n, \text{carrot} \rangle$	$\langle \text{eats}, ::n =d \text{ c}, \lambda x. \lambda y. \text{eat}(x)(y) \rangle$
$\langle \text{eat}, ::n =d \text{ c}, \lambda x. \lambda y. \text{eat}(x)(y) \rangle$	

Upon the latter revision, the agent may successfully derive rat , rats , and mouse , but also mouses , which will be rejected by the teacher. In order to avoid punishment, the learner had to wait for the well-formed item mice once to be uttered by \mathcal{T} . Yet, the current evidence prevents the agent from correctly segmenting eats , because our shallow semantic model does not sufficiently constrain any further pattern matching. This could possibly be remedied in case of sophisticated numeral and temporal semantic models. At the end of the day, we would expect something alike the hand-crafted lexicon (**Table 2**) above. For now, however, we leave this important problem for future research.

7. DISCUSSION

With the present study we have continued our work on language acquisition and on the unified description of physical interaction and linguistic communication of cognitive agents (Römer et al., 2019; beim Graben et al., 2020). The requirements for a unified description are primarily given by cognitive architectures. That includes the availability of suitable *representational data structures*, the satisfiability of the *composition-, adaptation- and classification principles* as well as the capability for *logical reasoning and learning*. Such an architecture will be particularly useful if it can explain the behavior of cognitive agents as well as the phylogenetic and ontogenetic development of language from earlier stages of evolution without language. Hence, the agent should be based on a physical symbol system (PSS) (Newell and Simon, 1976), that takes physical symbols from its sensory equipment, composing them into symbolic structures (expressions) and transforms them to new expressions that can generate goal directed actions. For this purpose we devised a grammar-based transformation mechanism that unifies physical interaction and linguistic communication using minimalist grammars (MG) and lambda calculus. To explain how the mechanism works, we have selected some example scenarios of the well known mouse-maze problem (Shannon, 1953). With this mechanism, the incoming sensory information from both cognitive loops can be brought together and transformed into meaningful information that can be saved now in a knowledge base and processed logically. The truth-functional approach that is required for the acquisition of a veridical model of a shared environment is dependent on this mechanism. Further, based on the uniform linguistic processing of interaction and communication the communication participants are able to exchange and synchronize their ideas about a common environment. The

use of the information arising from both cognitive loops is also useful here, since linguistic ambiguities can be resolved in this way.

Additionally, we have outlined an algorithm for effectively learning the syntax and semantics of English declarative sentences. Such sentences are presented to a cognitive agent by a teacher in form of utterance meaning pairs (UMP) where the meanings are encoded as formulas of first order predicate logic. This representation allows for the application of compositional semantics via lambda calculus (Church, 1936). For the description of syntactic categories we use Stabler's minimalist grammar (Stabler, 1997; Stabler and Keenan, 2003), (MG) a powerful computational implementation of Chomsky's recent Minimalist Program for generative linguistics (Chomsky, 1995). Despite the controversy between Chomsky and Skinner (Chomsky, 1995), we exploit reinforcement learning (Skinner, 2015; Sutton and Barto, 2018) as training paradigm. Since MG codifies universal linguistic competence through the five inference rules (10–11), thereby separating innate linguistic knowledge from the contingently acquired lexicon, our approach could potentially unify generative grammar and reinforcement learning, hence resolving the abovementioned dispute.

Minimalist grammar can be learned from linguistic dependency structures (Kobele et al., 2002; Stabler et al., 2003; Klein and Manning, 2004; Boston et al., 2010) by positive examples, which is supported by psycholinguistic findings on early human language acquisition (Pinker, 1995; Ellis, 2006; Tomasello, 2006). However, as Pinker (1995) has emphasized, learning through positive examples alone, could lead to undesired overgeneralization. Therefore, reinforcement learning that might play a role in children language acquisition as well (Moerk, 1983; Sundberg et al., 1996), could effectively avoid such problems. The required dependency structures are directly provided by the semantics in the training UMPs. Thus, our approach is explicitly semantically driven, in contrast to the algorithm of

Klein and Manning (2004) that regards dependencies as latent variables for EM training.

As a proof-of-concept we suggested an algorithm for simple English declarative sentences. We also have evidence that it works for German and French as well and hopefully for other languages also. Our approach will open up an entirely new avenue for the further development of speech-controlled cognitive user interfaces (Young, 2010; Baranyi et al., 2015).

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work was partly funded by the Federal Ministry of Education and Research in Germany under the grant no. 03IHS022A.

ACKNOWLEDGMENTS

This work was partly based on former joint publications of the authors with Werner Meyer, Günther Wirsching, and Ingo Schmitt (Wolff et al., 2018; beim Graben et al., 2019b, 2020; Römer et al., 2019). The content of this manuscript has been presented in part at the 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2019, Naples, Italy) (beim Graben et al., 2019b; Römer et al., 2019). According to the IEEE author guidelines a preprint of beim Graben et al. (2020) has been uploaded to arXiv.org.

REFERENCES

- Artzi, Y., and Zettlemoyer, L. (2013). Weakly supervised learning of semantic parsers for mapping instructions to actions. *Trans. Assoc. Comput. Linguist.* 1, 49–62. doi: 10.1162/tacl_a_00209
- Baranyi, P., Csapo, A., and Sallai, G. (2015). *Cognitive Infocommunications (CogInfoCom)*. Springer. ISBN 9783319196077.
- beim Graben, P., Meyer, W., Römer, R., and Wolff, M. (2019a). "Bidirektionale Utterance-Meaning-Transducer für Zahlwörter durch kompositionale minimalistische Grammatiken," in *Tagungsband der 30. Konferenz Elektronische Sprachsignalverarbeitung (ESSV), Volume 91 of Studientexte zur Sprachkommunikation*, eds P. Birkholz and S. Stone (Dresden: TU-Dresden Press), 76–82.
- beim Graben, P., Römer, R., Meyer, W., Huber, M., and Wolff, M. (2019b). "Reinforcement learning of minimalist numeral grammar," in *Proceedings of the 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Naples: IEEE), 67–72.
- beim Graben, P., Römer, R., Meyer, W., Huber, M., and Wolff, M. (2020). Reinforcement learning of minimalist grammars. *arXiv[Preprint].arXiv:2005.00359*.
- Bischof, N. (2009). *Psychologie, ein Grundkurs für Anspruchsvolle*, 2nd Edn. Stuttgart: Verlag Kohlhammer.
- Bonato, R., and Retoré, C. (2001). "Learning rigid Lambek grammars and minimalist grammars from structured sentences," in *Proceedings of the Third Workshop on Learning Language in Logic (Strasbourg)*, 23–34.
- Boston, M. F., Hale, J. T., and Kuhlmann, M. (2010). "Dependency structures derived from minimalist grammars," in *The Mathematics of Language, Vol. 6149 of Lecture Notes in Computer Science*, eds C. Ebert, G. Jäger and J. Michaelis (Berlin: Springer), 1–12.
- Boullier, P. (2005). "Range concatenation grammars," in *New Developments in Parsing Technology, volume 23 of Text, Speech and Language Technology*, eds H. Bunt, J. Carroll, and G. Satta (Berlin: Springer), 269–289.
- Chen, P. P. S. (1975). "The entity-relationship model: toward a unified view of data," in *Proceedings of the 1st International Conference on Very Large Data Bases (VLDB75)*, ed D. S. Kerr (New York, NY: ACM Press), 173.
- Chomsky, N. (1995). *The Minimalist Program. Number 28 in Current Studies in Linguistics*, 3rd printing 1997. Cambridge, MA: MIT Press.
- Church, A. (1936). An unsolvable problem of elementary number theory. *Am. J. Math.* 58, 345. doi: 10.2307/2371045

- Diessel, H. (2013). "Construction grammar and first language acquisition," in *The Oxford Handbook of Construction Grammar*, eds T. Hoffmann and G. Trousdale (Oxford: Oxford University Press), 347–364.
- Eliasmith, C. (2013). *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford: Oxford University Press.
- Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Appl. Linguist.* 27, 1–24. doi: 10.1093/applin/ami038
- Funke, J. (ed.). (2006). *Denken und Problemlösen, Enzyklopädie der Psychologie, 8th Edn*. Hogrefe: Verlag für Psychologie.
- Gee, J. P. (1994). First language acquisition as a guide for theories of learning and pedagogy. *Linguist. Educ.* 6, 331–354. doi: 10.1016/0898-5898(94)90002-7
- Gold, E. M. (1967). Language identification in the limit. *Inform. Control* 10, 447–474. doi: 10.1016/S0019-9958(67)91165-5
- Haegeman, L. (1994). *Introduction to Government Binding Theory, volume 1 of Blackwell Textbooks in Linguistics, 2nd Edn, 1st Edn 1991*. Oxford: Blackwell Publishers.
- Hale, J. T. (2011). What a rational parser would do. *Cogn. Sci.* 35, 399–443. doi: 10.1111/j.1551-6709.2010.01145.x
- Harkema, H. (2001). *Parsing Minimalist Languages* (Ph.D. thesis), University of California, Los Angeles, CA.
- Harnad, S. (1990). The symbol grounding problem. *Physica D* 42, 335–346. doi: 10.1016/0167-2789(90)90087-6
- Hausser, R. (2014). *Foundations of Computational Linguistics*. Berlin; Heidelberg: New York, NY: Springer Verlag.
- Joshi, A. K., Levy, L. S., and Takahashi, M. (1975). Tree adjunct grammars. *J. Comput. Syst. Sci.* 10, 136–163. doi: 10.1016/S0022-0000(75)80019-5
- Joshi, A. K., Vijay-Shanker, K., and Weir, D. (1990). *The convergence of mildly context-sensitive grammar formalisms*. Technical Report MS-CIS-90-01, University of Pennsylvania.
- Jungclaussen, H. (2001). *Kausale Informatik*. Wiesbaden: Deutscher Universitätsverlag.
- Klein, D., and Manning, C. D. (2004). "Corpus-based induction of syntactic structure: models of dependency and constituency," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, eds D. Scott, W. Daelemans, and M. A. Walker (Stroudsburg, PA: Association for Computational Linguistics), 478–485.
- Kobele, G. M. (2009). "Syntax and semantics in minimalist grammars," in *Proceedings of ESSLLI 2009* (Bordeaux).
- Kobele, G. M., Collier, T. C., Taylor, C. E., and Stabler, E. P. (2002). "Learning mirror theory," in *Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG+2002)*, ed R. Frank (Stroudsburg, PA: Association for Computational Linguistics), 66–73.
- Kracht, M. (2003). *The Mathematics of Language*. Berlin; New York, NY: Mouton de Gruyter.
- Kuhlmann, M., Koller, A., and Satta, G. (2015). Lexicalization and generative power in CCG. *Comput. Linguist.* 41, 215–247. doi: 10.1162/COLI_a_00219
- Kwiatkowski, T., Goldwater, S., Zettlemoyer, L., and Steedman, M. (2012). "A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL '12* (Stroudsburg, PA: Association for Computational Linguistics), 234–244.
- Lausen, G. (2005). *Datenbanken, Grundlagen und XML-Technologien*. München: Elsevier GmbH, Spektrum Akademischer Verlag. 1. Auflage.
- Lohnstein, H. (2011). *Formale Semantik und natürliche Sprache, 2nd Edn*. Berlin: De Gruyter.
- Mainguy, T. (2010). A probabilistic top-down parser for minimalist grammars. *ArXiv, abs/1010.1826v11*.
- Michaelis, J. (2001). "Derivational minimalism is mildly context-sensitive," in *Logical Aspects of Computational Linguistics, Volume 2014 of Lecture Notes in Artificial Intelligence* (Berlin: Springer), 179–198.
- Moerk, E. L. (1983). A behavioral analysis of controversial topics in first language acquisition: reinforcements, corrections, modeling, input frequencies, and the three-term contingency pattern. *J. Psycholinguist. Res.* 12, 129–155. doi: 10.1007/BF01067408
- Newell, A., and Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Commun. ACM* 19, 113–126. doi: 10.1145/360018.360022
- Nivre, J. (2003). "An efficient algorithm for projective dependency parsing," in *Proceedings of the Eighth International Workshop on Parsing Technologies (IWPT2003)* (Nancy), 149–160.
- Niyogi, S. (2001). "A minimalist implementation of verb subcategorization," in *Proceedings of the Seventh International Workshop on Parsing Technologies (IWPT2001)* (Beijing).
- Pinker, S. (1995). "Language acquisition," in *Language: An Invitation to Cognitive Science, Chapter 6*, eds L. R. Gleitman, D. N. Osherson, M. Liberman, L. R. Gleitman, D. N. Osherson, and M. Liberman (Cambridge, MA: MIT Press), 135–182.
- Römer, R., beim Graben, P., Huber, M., Wolff, M., Wirsching, G., and Schmitt, I. (2019). "Behavioral control of cognitive agents using database semantics and minimalist grammars," in *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Naples).
- Schönfinkel, M. (1924). Über die Bausteine der mathematischen Logik. *Mathematische Annalen* 92, 305–316. doi: 10.1007/BF01448013
- Seki, H., Matsumura, T., Fujii, M., and Kasami, T. (1991). On multiple context-free grammars. *Theor. Comput. Sci.* 88, 191–229. doi: 10.1016/0304-3975(91)90374-B
- Shannon, C. E. (1953). Computers and automata. *Proc. Institute Radio Eng.* 41, 1234–1241. doi: 10.1109/JRPROC.1953.274273
- Skinner, B. F. (2015). *Verbal Behavior, 1st Edn 1957*. Mansfield Centre, CT: Martino Publishing.
- Stabler, E. (1997). "Derivational minimalism," in *Logical Aspects of Computational Linguistics (LACL96), volume 1328 of Lecture Notes in Computer Science*, ed C. Retoré (New York, NY: Springer), 68–95.
- Stabler, E. P. (2011a). "Computational perspectives on minimalism," in *Oxford Handbook of Linguistic Minimalism*, ed C. Boeckx (Oxford: Oxford University Press), 617–641.
- Stabler, E. P. (2011b). "Top-down recognizers for MCFGs and MGs," in *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics* (Portland, OR: Association for Computational Linguistics), 39–48.
- Stabler, E. P., Collier, T. C., Kobele, G. M., Lee, Y., Lin, Y., Riggle, J., et al. (2003). "The learning and emergence of mildly context sensitive languages," in *Advances in Artificial Life, volume 2801 of Lecture Notes in Computer Science*, eds W. Banzhaf, T. Christaller, P. Dittrich, J. T. Kim, and J. Ziegler (Berlin: Springer), 525–534.
- Stabler, E. P., and Keenan, E. L. (2003). Structural similarity within and among languages. *Theor. Comput. Sci.* 293, 345–363. doi: 10.1016/S0304-3975(01)00351-6
- Stanojević, M., and Stabler, E. (2018). "A sound and complete left-corner parsing for minimalist grammars," in *Proceedings of the Eight Workshop on Cognitive Aspects of Computational Language Learning and Processing*, eds M. Idiat, A. Lenci, T. Poibeau, and A. Villavicencio (Stroudsburg, PA: Association for Computational Linguistics), 65–74.
- Sundberg, M. L., Michael, J., Partington, J. W., and Sundberg, C. A. (1996). The role of automatic reinforcement in early language acquisition. *Anal. Verbal Behav.* 13, 21–37. doi: 10.1007/BF03392904
- Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA; London: The MIT Press.
- Tomasello, M. (2006). First steps toward a usage-based theory of language acquisition. *Cogn. Linguist.* 11:61. doi: 10.1515/cogl.2001.012
- van Zaanen, M. (2001). "Bootstrapping syntax and recursion using alignment-based learning," in *Proceedings of the Seventeenth International Conference on Machine Learning* (San Francisco, CA), 1063–1070.
- Vera, A., and Simon, H. (1993). Situated action: a symbolic interpretation. *Cogn. Sci.* 17, 7–48. doi: 10.1207/s15516709cog1701_2
- Wegner, P. (2003). *Lambda Calculus*. Hoboken, NJ: John Wiley and Sons Ltd., GBR.
- Wirsching, G., and Lorenz, R. (2013). "Towards meaning-oriented language modeling," in *Proceedings of the 4th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest), 369–374.
- Wolff, M., Huber, M., Wirsching, G., Römer, R., Graben, P., Schmitt, I. (2018). "Towards a quantum mechanical model of the inner stage of cognitive agents," in *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Budapest: IEEE).
- Wolff, M., Römer, R., and Wirsching, G. (2015). "Towards coping and imagination for cognitive agents," in *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Gyor: IEEE), 307–312.
- Wolff, M., Tschpe, C., Römer, R., and Wirsching, G. (2013). "Subsymbol-symbol-transduktoren," in *Proceedings of "Elektronische Sprachsignalverarbeitung"*

- (ESSV),” volume 65 of *Studententexte zur Sprachkommunikation*, eds P. Wagner (Dresden: TUDpress), 197–204.
- Young, S. (2010). “Still talking to machines (cognitively speaking),” in *Conference of the International Speech Communication Association, INTERSPEECH 2010*, eds T. Kobayashi, K. Hirose, and S. Nakamura (Makuhari: ISCA), 1–10.
- Zettlemoyer, L. S., and Collins, M. (2005). “Learning to map sentences to logical form: structured classification with probabilistic categorial grammars,” in *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence, UAI’05* (Arlington, VA: AUAI Press), 658–666.

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.

Copyright © 2022 Römer, beim Graben, Huber-Liebl and Wolff. 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.

Frontiers in Computer Science

Explores fundamental and applied computer science to advance our understanding of the digital era

An innovative journal that fosters interdisciplinary research within computational sciences and explores the application of computer science in other research domains.

Discover the latest Research Topics

[See more →](#)

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact

