Experientially grounded language production:

Advancing our understanding of semantic processing during lexical selection

Dissertation

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.) im Fach Psychologie

eingereicht an der Lebenswissenschaftlichen Fakultät der
Humboldt-Universität zu Berlin

von
Anne Vogt, geb. Lohse

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr. Julia von Blumenthal

Dekan der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin
Prof. Dr. Dr. Christian Ulrichs

Gutachter/innen:

1. Prof. Dr. Rasha Abdel Rahman
2. Prof. Dr. Berry Claus
3. Dr. Kristof Strijkers

Eingereicht am: 27.10.2022

Mündliche Prüfung am: 13.02.2023

Semantics has been, and still is, a surprisingly impractical occupation.

Magnus Sahlgren

Aus Keplers wunderbarem Lebenswerk erkennen wir besonders schön, daß aus bloßer Empirie allein die Erkenntnis nicht erblühen kann, sondern aus dem Vergleich des Gedachten mit dem Beobachteten.

Albert Einstein

Table of Contents

Acknowledgements

> It takes a village to raise a child.
>
> *African proverb*

This third child of mine indeed would never have been raised to a grown-up state were it not for the help and support of many, many people to whom I am deeply grateful.

First and foremost, thank you EMANUEL for your love, patience, guidance, and understanding and for taking away as many everyday life pressures from me as you could. Thanks for listening to all-things-language whenever I became enthusiastic during all these years and for sometimes pushing me out of my comfort zone. I am also incredibly thankful that you two, MORITZ and PAULA, reminded me once in a while what is really important in life. Especially, during the months we all spent at home due to Corona you were incredibly supportive, responsible and understanding and I am very happy how we all went through that time. Your joy, creativity and affection always brought back my energy.

I am also very thankful to my wider family: my parents, my parents in law and my sister with her family. You were there whenever help and support was needed.

Without the help and guidance of my supervisor RASHA ABDEL RAHMAN this dissertation would not have seen the light of the day. Thank you for challenging my views in discussions, for giving me feedback, for developing my research ideas and last but not least for gathering such a fine group of Neuros around you. I would not have ended up in the Neuro lab had it not been for BARBARA KAUP. Thank you, for telling me once that sometimes researchers simply need luck. I am also very thankful for the trust you placed in me when I started in your group after my studies. Furthermore, your example as a mother of three doing research has inspired me a lot. I am also grateful to MARTIN FISCHER and his group at the University of Potsdam (especially ARTIANNA FELISATTI as well as ALEX MIKLASHEVSKY) who saw this work grow from its first stages and always provided helpful feedback whenever I presented it to them.

Luckily, I was not the only one raising such a child of science and I found great support, help and also friendship among my Neuro colleagues and my colleagues from Tübingen. Anna Eiserbeck, thanks for always answering my questions or for listening to me whenever I was turning around from my desk. Thanks, Roger Hauber for fighting the pandemia together with me. This online experiment project opened up my eyes that I need to work in teams rather than alone and I am very happy that we took up this challenge together and managed to make it such a success. Kirsten Stark, thank you for getting inspired and sharing my enthusiasm for things not directly related to our research which led to a file drawer project for the Studienstiftung. I greatly enjoyed spending time with you at the ESCOP in Lille. Anna-Lisa Döring, you were my soulmate in these last months. Talking to you was both highly motivating and a true comfort. Thank you for being there and for sharing all the ups and downs. Thanks to you and to Antje Lorenz for your valuable feedback on earlier drafts of this dissertation. I am very thankful to Annika Just and Nura Völk for all the chitdchat about non-research items and for your emotional support and coaching about inclusive language. Carolin Dudschig helped me a lot by providing me the opportunity to work with children and to write my very first paper - I learned a lot from this experience which helped me in later phases of my PhD. Luckily, Fritz Günther never refused to answer my questions. Sharing some of my confusion about meaning, representations and the lexicon with you helped me to deal with the uncertainty that comes when we are becoming explorers at the edge of the unknown. Thanks to all the other Neuros, too, for coffee, answering questions, sharing code, solving technical problems, talking about all or nothing and for always supporting each other: Hannah Kaube, Anna Kuhlen, Martin Maier, Julia Baum, Alex Enge, Hsin Pei Lin, Martina Erhardt, Nora Holtz, Haluk Öngören and Guido Kiecker. Some of you were also involved in my research as student assistants and I would not have been able to complete it in the given time were it not for your skills, time and effort.

Doing my research I also greatly profited from all the students I supervised. Not only did you take up tasks like testing participants and preprocessing huge amounts of audio files. You also taught me how to speak about my research in an intelligible way. Talking to you always reignited my interest in the subject matter. Working with you I realized that I love to teach and explain and that I care a lot for people and their development. I am most thankful to Tabea Villinger who was the first student I supervised and with whom I had very deep and philosophical discussions over embodiment. You gave me the time I needed to become the supervisor I wanted to be. I am also very grateful to Isabel Ganter who had so many ideas and an incredible amount of independence which made supervising so easy and the project we developped together such a cool one.

In times of doubt I sought and found help thanks to several senior people who mentored me and helped me to find my place, both in the world of research as well as beyond. Thank you, Petra Metz for you personal support and for the opportunities you opened up with the Women In Natural Sciences Adlershof program. You introduced me to Helga Schwalms and I am glad you shared your personal experiences so openly. This helped me to find my new role as a working mum and researcher and to settle in Berlin.

Thanks to the Karriereförderprogramm für Frauen der Begabtenförderungswerke my self esteem got a big boost and I received lots of power and energy for the last part of my PhD from this wonderful network. I was always looking forward to our meetings, learned so much from you and will always remember some of your words, Anna-Katharina Schak. Being accompanied in this rough journey of finishing a PhD, managing a family, and trying to set foot outside of academia in a very warm and yet professional way was a true gift! With regard to this dissertation, Laura Voigt helped me to set my priorities and to gain focus for finalizing this project.

I am very thankful for the zoom coffee meetings during which we shared the struggles in academia as a mother during Corona, Jo Brüggemann and Maria Velte. Thank you, Leonie Föbus, for making me realize when my eyes are shining when talking about my job perspectives. Thank you, Özgür Kesim for adapting to our rhythm and for coming to our place so often to make music, drink wine and talk about life. You also helped me to understand the audio recording code and to make it a little smoother when we implemented the online language production experiment. Very important, too, were my dear friends Carleen Kluger and Katharina Krause. Thanks for keeping the connection alive even though physical distance, a family life and our ambitions lead to tight schedules for all of us. You reminded me of my strengths and also of the fact that being a working mum should never let us become like the armed warrior goddess Athena.

Thanks also go to the Berlin School of Mind and Brain where I met many young and inspiring researchers, learned new skills and received funding for conferences and to the Stiftung der Deutschen Wirtschaft who awarded me a PhD scholarship before I became a sholarship holder from the Studienstiftung des deutschen Volkes for four years. I am especially thankful for their support from the beginning of my studies onwards which helped me to follow my interest in language and cognition by studying two subjects. This was definitely not the fast track and therefore, I am even more thankful that the stipend allowed me to follow Humboldt's ideal. Not only did I receive money, I also gathered some of my dearest memories of my study and PhD time taking part in the program organized by the Studienstiftung.

Zusammenfassung

Wenn wir sprechen, müssen wir die richtigen Wörter wählen, um die von uns intendierte Botschaft zu übermitteln und erfolgreich zu kommunizieren. Dieser Prozess der lexikalischen Auswahl ist bislang nicht hinreichend verstanden. Insbesondere wurde kaum erforscht, inwiefern Bedeutungsaspekte, welche in sensomotorischen Erfahrungen gründen, diesen Prozess der Sprachproduktion beeinflussen. Die Rolle dieser Bedeutungsaspekte wurde mit zwei Studien untersucht, in denen Probanden Sätze vervollständigten. In Studie 1 wurde der visuelle Eindruck der Satzfragmente manipuliert, so dass die Sätze auf- oder absteigend am Bildschirm erschienen. In Studie 2 mussten die Probanden Kopfbewegungen nach oben oder unten ausführen, während sie die Satzfragmente hörten. Wir untersuchten, ob räumliche Aspekte der produzierten Wörter durch die räumlichen Manipulationen sowie die räumlichen Eigenschaften der präsentierten Satzfragmente beeinflusst werden. Die vertikale visuelle Manipulation in Studie 1 wirkte sich nicht auf die räumlichen Attribute der produzierten Wörter aus. Die Kopfbewegungen in Studie 2 führten zu einem solchen Effekt – Kopfbewegungen nach oben führten dazu, dass die Referenten der produzierten Wörter weiter oben im Raum angesiedelt waren als nach Bewegungen nach unten (und anders herum). Darüber hinaus war dieser Effekt stärker, je ausgeprägter die interozeptive Sensibilität der Probanden war, d.h. je stärker sie Körpersignale wahrnehmen und sich darauf konzentrieren konnten. Außerdem beeinflussten die räumlichen Aspekte der Satzfragmente die räumlichen Eigenschaften der produzierten Wörter in beiden Studien. Somit zeigt diese Arbeit, dass in der Erfahrung basierende Bedeutungsanteile, welche entweder in Sprache eingebettet sind oder durch körperliche Aktivität reaktiviert werden, die Auswahl der Wörter beim Sprechen beeinflussen und dass interindividuelle Unterschiede diesen Effekt modulieren können. Die Befunde werden in Bezug zu Theorien der Semantik gesetzt.

Darüber hinaus wird mit dieser Dissertation das Methodenrepertoire von Sprachproduktionsforschern erweitert. Studie 3 bietet einen methodischen Ansatz für die Durchführung von Online-Sprachproduktionsexperimenten mit Bildbenennung.

Abstract

For communication to be successful, we have to produce the right words to get the intended message across. This process of lexical selection is not well understood. Specifically, meaning aspects grounded in sensorimotor experiences and their role during lexical selection have not been investigated widely in language production. To close this gap, we investigated the role of experientially grounded meaning aspects with two studies in which participants had to produce a noun to complete sentences which described sceneries. In Study 1, the visual appearance of sentence fragments was manipulated and they seemed to move upwards or downwards on screen. In Study 2, participants moved their head up- or downwards while listening to sentence fragments. We investigated whether the spatial properties of the freely chosen nouns are influenced by the spatial manipulations as well as by the spatial properties of the sentences. The vertical visual manipulation used in Study 1 did not influence the spatial properties of the produced words. However, the body movements in Study 2 influenced participants' lexical choices, i.e. after up-movements the referents of the produced words were higher up compared to after downward movements (and vice verse). Furthermore, there was an increased effect of movement on the spatial properties of the produced nouns with higher levels of participants' interoceptive sensibility, i.e. sensitivity and attention towards bodily signals. Additionally, the spatial properties of the stimulus sentences influenced the spatial properties of the produced words in both studies. Thus, the present work shows that experientially grounded meaning aspects which are either embedded in text or reactivated via bodily manipulations may influence which words we chose when speaking, and interindividual differences may moderate these effects. The findings are related to current theories of semantics.

Furthermore, this dissertation enhances the methodological repertoire of language production researchers by providing a proof of concept for running language production studies with overt articulation in picture naming tasks online (Study 3).

1 Introduction

The use of language is distinguishing humans from all other species allowing us to communicate with our conspecifics across time and space. Despite being of key importance for human nature, we still do not really understand why and how we choose a certain word in order to convey an intended meaning. Moreover, it is unclear what constitutes the meaning of a word and how we access it. These question have already been raised more than 2000 years ago, yet they are still investigated today (Kiefer & Pulvermüller, 2012).

Within the framework of *Embodied Cognition* researchers and philosophers investigated in how far mental processes might be influenced by the form of our bodies as well as by our sensory and motor systems (Barsalou, 1999; Glenberg, 2010). Inspired by theoretical arguments that symbols need to be grounded in order to be meaningful (Harnad, 1990; Searle, 1980), the idea that our bodies play a pivotal role for processes like thinking and communicating became a topic for empirical investigations especially with regard to meaning processing in language. Researchers changed their focus from viewing language processing as being based on abstract, arbitrary, amodal symbols (Fodor, 1975) to language and meaning being grounded in sensorimotor experiences (Barsalou, 1999). While this has led to many debates and insights (e.g., Mahon & Caramazza, 2008; Meteyard, Cuadrado Rodriguez, Bahrami, & Vigliocco, 2012; Ostarek & Huettig, 2019; Pulvermüller, 2013) the empirical investigations of embodied meaning processing have neglected language production as most studies focused on language comprehension.

With this dissertation I want to close this gap by prodiving evidence that our lexical choices during language production are influenced by our bodily states and by the reactivation of experiential traces which are grounded in sensory experiences. In Study 1, a sentence completion task was used to investigate whether visual stimulations and experientially grounded meaning aspects of vertical space influence lexical choices. Adding head movement manipulations, Study 2 explored whether body posture changes impact the lexical choices in the same task, and in how far embodied language production depends on interindividual differences in interoceptive sensibility. Furthermore, I present an implementation of running language production experiments requiring overt responses on the internet in Study 3, a setup which was used for Study 2.

In the following sections the reader is provided with the theoretical foundations as well as empirical insights which build the basis for these studies. First, the theoretical framework of embodied cognition is introduced in Section 1.1.1, followed by relevant findings from language comprehension in Section 1.1.2 as most empirical research on embodied language processing has been conducted in language comprehension. A special focus is laid on evidence regarding language-space associations in Section 1.1.3. Next, psycholinguistic models of language production are introduced in Section 1.2.1 and evidence on semantic processing in language production is presented in Section 1.2.2. Empirical evidence for embodied semantic processing in language production is presented in more detail in Section 1.2.3. As empirical work on language production was hampered due to Covid-19, methodological challenges which had to be overcome during that time are discussed in Section 1.2.4. I end the Introduction with the aims of my dissertation, along with an outline of the present work in Section 1.3.

1.1 Embodied cognition and language processing

1.1.1 Theoretical aspects

Communicating about experiences we make in the world is a core function of human language. We hear or speak about certain objects, emotions, situations or actions we experience and as a consequence, specific words or any linguistic constructions frequently co-occur with the objects, situations and experiences we talk about. Memories of these situations or objects are laid down in multiple modality-specific cortical areas, e.g. because the specific object was related to something we could see, smell, touch and also hear when its corresponding name was uttered (Kiefer & Pulvermüller, 2012). Thus, these so-called multimodal experiential traces are linked to our knowledge of objects, situations and actions and also to the words or phrases used to refer to these things (e.g., Lynott, Connell, Brysbaert, Brand, & Carney, 2020; Zwaan & Madden, 2005). The unintentional und unconscious reactivation of these experiential traces when we read or hear language has been called experiential simulation (Pecher, van Dantzig, & Schifferstein, 2009; Zwaan & Madden, 2005).

Due to the necessarily individual character of experiences, the embodied cognition framework predicts differential effects in cognitive processing which might hinge on levels of expertise (e.g., Holt & Beilock, 2006; Wolter, Dudschig, & Kaup, 2017), different bodily dispositions (e.g., Klostermann, Wyrobnik, Boll, Ehlen, & Tiedt, 2022; Meteyard et al., 2012), individual processing preferences (e.g., Yee & Thompson-Schill, 2016) but also on different levels of sensitivity towards bodily signals (Häfner, 2013; Herbert & Pollatos, 2012; Villani, Lugli, Liuzza, Nicoletti, & Borghi, 2021).

The building blocks of our semantic knowledge have often been called <concepts>. They are a mental symbol to describe the cognitive units forming the base of our knowledge (see the Cognitive Atlas for this definition of <concept>, Poldrack et al., 2011). From an embodied perspective, concepts are viewed as experience-dependent and flexible representations which are modality specific. Neurally, these representations are distributed in networks including sensorimotor areas of the brain, without being restricted to these (e.g., Lambon Ralph, 2014; Pulvermüller, 2018). Accessing the concepts with the purpose of retrieving a word's meaning leads to the partial reactivation of the same brain processes which are active whenever we experience the things or actions to which the word refers. This reuse of sensorimotor brain areas in a context-dependent way enables meaning construction (e.g., Pulvermüller, 2018). Thus, these reactivations are not mere by-products of language processing due to our learning history but are thought of as constructive processes. However, the degree to which they are functionally relevant is still a matter of debate (e.g., Ostarek & Bottini, 2020).

The focus of empirical investigations of embodied language processing has been mainly on those meaning aspects which are linked to sensorimotor experiences. However, it has been stated that language itself can also be experienced and thus, become one of the sources for grounding meaning. Thus, the way words are used during communication and in texts forms our linguistic experience, which is one of the sources from which we acquire semantic knowledge (Günther, Rinaldi, & Marelli, 2019; Louwerse, 2018; Vinson, Andrews, & Vigliocco, 2014). This view has been reflected in distributional models of semantics. Because words which are more similar in meaning tend to occur in similar linguistic contexts, the meaning of a word can be inferred from its statistical distribution in spoken and written language (Firth, 1957; Sahlgren, 2008). Accordingly, distributional models quantify the co-occurrence of words across texts and contexts and thereby provide a quantitative measure of meaning similarity (for an overview see Günther et al., 2019).

In this thesis, the phrase *embodied* or *sensorimotor* meaning will be used whenever referring to meaning aspects which are closely linked to sensorimotor experiences we can make with our bodies in the world, e.g. the visual appearance, smell, taste, or haptic related to a fruit which we label with the word *apple.* The phrase *experientially grounded* meaning will be used in a broader sense, encompassing both embodied meaning as well as meaning aspects which are inferred from language usage, e.g. that apples are forbidden fruits.

1.1.2 Evidence for embodiment in language comprehension

I will now present exemplary evidence supporting the above theory of embodied language comprehension. In some cases alternative, non-embodied explanations for the presented findings may exist but I will mostly present the interpretation given by the original authors. Thereby, the reader is provided with an overview over both relevant findings as well as methodologies used to investigate embodied language comprehension. In Section 1.1.3 I will focus on empirical language comprehension studies which are more directly related to the experiments conducted within this dissertation.

Empirical investigations have been focusing on the interaction of language with perception or action and much evidence supports the view that language comprehension is embodied to some extent (for overviews see e.g., Bergen, 2015; Meteyard et al., 2012). In most experiments, participants typically process some linguistic input and subsequent modulations of motor actions or performance in perceptual tasks are measured or neurophysiological measures are being recorded (Kaup, de la Vega, Strozyk, & Dudschig, 2016).

For example, participants had to indicate whether a depicted object was mentioned in the previous sentence (Zwaan, Stanfield, & Yaxley, 2002). The descriptions varied so that the implied shape of an object depended on the respective description, e.g., a roundish shape for the sentence <*The egg is in the pan*> describing supposedly a fried egg versus an oval shape for the sentence <*The egg is in the fridge*>. Participants were faster to respond to the probe question if the shape implied by the sentence matched the shape in the presented drawing. Thus, participants seem to automatically infer the shape of things via visual simulations when they are mentioned (for an alternative explanation see Ostarek & Bottini, 2020).

Using pupillary responses as dependent measure, it was found that participants' pupil size was predicted by the sense of brightness or darkness conveyed by words presented to them (Mathôt, Grainger, & Strijkers, 2017). Pupils were larger for words conveying a sense of darkness (e.g., *shade*) compared to words conveying a sense of brightness (e.g., *lamp*). This finding suggests that the meaning of a word is sufficient to trigger bodily responses even when the meaning dimension leading to the bodily response is beyond voluntary control and not relevant for the experimental task (categorization of words as referring to animals vs. not).

Evidence for the reactivation of sensorimotor brain areas during language processing stems from functional magnetic resonance imaging (fMRI) studies. In a seminal experiment, participants had to passively read action words referring to face, arm or leg related actions (Hauk, Johnsrude, & Pulvermüller, 2004). When participants read verbs like *lick*, *pick* or *kick* the brain was activated in the motor cortex and adjacent areas. These areas correspond closely to the somatotopic map of the body parts involved in those actions.

Similarly, processing odour related words like *vanilla* leads to increased activation in the primary olfactory cortext (González et al., 2006) and processing taste related words like *salt* leads to increased activation in the primary and secondary gustatory cortex (Barrós-Loscertales et al., 2012). However, the processing of language related to smell, taste and touch seems to rely less on sensorimotor simulation compared to language with visual, motor-related or auditory content (Speed & Majid, 2019).

For concepts that are associated with specific sounds, such as <telephone> or <hair dryer>, partial reactivation of brain areas linked to auditory processing has been demonstrated using a lexical decision task, both with fMRI and electroencephalography (EEG) (Kiefer, Sim, Herrnberger, Grothe, & Hoenig, 2008). The auditory relevance of probe-words was manipulated. Compared to words for which auditory features are less relevant but which were carefully matched for other psycholinguistic variables, words like *telephone* led to increased activation in auditory brain areas starting from 150 ms post stimulus onset. Converging evidence for this early activation of task-irrelevant experientially grounded semantic features comes from a study using magnetoencephalography (MEG) (García et al., 2019). It was shown that action words started to modulate activity in motor areas from around 130 ms after word presentation which happened earlier than in non-modal brain areas which have traditionally been linked to meaning processing like the anterior temporal lobe (ATL). The rapidity of these activations indicates that the sensorimotor reactivations are not mere by-products of our history of experiences, which only show up when the meaning computation has already happened. Instead, the reactivations seem to play a crucial role for meaning construction and understanding. This idea is corroborated by patient studies showing that lesions to modality specific or motor brain areas result in selective deficits in processing words with sensory or motor content (e.g., Dreyer et al., 2015; Trumpp, Kliese, Hoenig, Haarmeier, & Kiefer, 2013).

With the Lancaster Sensorimotor Norms data base, a composite score of sensory strength has been introduced which is a mathematical computation yielding one measure to describe the degree to which concepts are grounded in the senses (Lynott et al., 2020). Concepts can have similar sensory strength even though being grounded in different modalities. For example, <rain> and <lemon> have similar sensory strength values according to the data base. It has been demonstrated that this composite score is an important predictor in psycholinguistic tasks like lexical decision (Lynott et al., 2020). Furthermore, this score of perceptual strength also encompasses a dimension which is linked to interoception, that is our ability to identify, access, understand and react towards internal bodily signals (Craig, 2002; Garfinkel, Seth, Barrett, Suzuki, & Critchley, 2015). Interoceptive experiences seems to be highly relevant for grounding emotion concepts but also more abstract concepts (e.g., Connell, Lynott, & Banks, 2018; Villani et al., 2021).

This short overview of empirical evidence suggests that concepts are grounded in sensorimotor experiences, and that they fulfil the criteria for grounded meaning set up by Kiefer et al. (2008): conceptual processing during implicit tasks automatically activates perceptual features and sensorimotor brain regions rapidly and in a selective manner. The above mentioned studies also show that sensorimotor reactivations during reading or listening can be found in a variety of tasks using different methodologies. This suggests that the observed effects cannot solely be attributed to the methodology used. Furthermore, they are found from early age on (Wellsby & Pexman, 2014).

However, there is also evidence that does not fit well with an account of embodied language processing. For example, Glenberg and Kaschak (2002) reported the action-sentence-compability effect, which became a hallmark finding of embodied language comprehension. Participants had to press a button to judge the sensibility of sentences with an implied direction towards or away from the body like *<You gave Liz the book>* or *<You open the drawer>*. The button presses either also required a movement away from the body or towards it. If there was a match between the direction implied by the sentence and the movement direction for pressing the button, participants reacted faster. These findings were taken as evidence that understanding language is grounded in bodily actions. However, the overall effect size of the action-sentence-compatibility effect is small and the effect does not replicate well (see e.g., Morey et al., 2022; Papesh, 2015; A. Winter, Dudschig, Miller, Ulrich, & Kaup, 2022). Moreover, in similar setups relying on the compatibility of sentence content and participants' responses, both facilitatory and interfering effects have been taken as evidence for embodied language comprehension (compare the results by Connell, 2007 and Zwaan & Pecher, 2012). Whether compatibility paradigms lead to facilitation or interference seems to be context- and task-dependent (for a discussion see e.g., Gozli, Chasteen, & Pratt, 2013). Furthermore, it has been shown that a lexical decision task with hand- and foot-related words was not influenced by concurrent hand or foot movements. If perceptual simulation would be necessary for lexical decision, an interfering effect of concurrent movement would have been expected. Therefore, it has been argued that experiential simulations might not be functionally relevant to understand language – at least not in a lexical decision task (Strozyk, Dudschig, & Kaup, 2019).

Due to the inconsistency of findings and the unclear relevance of experiential simulation, the question whether language processing is built on mental simulations of previous experiences is not fully answered. Ostarek and Huettig (2019) list several challenges which need to be addressed by the research community in the future to answer this question: (1) Decisive paradims need to be developped, (2) causality needs to be probed, (3) task dependency of embodied language processing needs to be understood, (4) explicit predictions about the direction and timing of effects are needed, (5) emerging theories should be assessed with novel neuroimaging methods and (6) an all-encompassing theory is needed. One more challenge should be added, namely to investigate the role of experiential simulation in language production.

In this dissertation, I will specifically tackle one of the above challenges by contributing to the following research question with Study 1 and 2: Are experiential reactivations causally relevant for language processing?

1.1.3 Reactivation of experiential traces of vertical space

The experiential domain of vertical space has been one of the meaning aspects which received much attention in the embodied language processing literature. It lends itself well as a test case because the knowledge of where in space certain objects and things typically occur has traditionally not been considered a meaning feature. To assess participants' knowledge of the semantic content of words, feature listing tasks have often been used (McRae, Cree, Seidenberg, & McNorgan, 2005; Speed, Vinson, & Vigliocco, 2015). Although the typical spatial location of a word's referent can be easily inferred, people usually do not list features like DOWN for a word like *puddle* or UP for a word like *bird's nest*. A second advantage of investigating vertical space is that objects typically sharing the same space in the lower or upper sphere do not form a natural category or are necessarily thematically or associatively linked to each other (consider for example *shoe*, *rail track* and *mushroom*). I use the term *language-space associations* for this mutual knowledge of where we typically find the referents of these concepts (see also e.g., Dudschig & Kaup, 2017). Note that the spatial semantics of the word class spatial prepositions like *in/out*, *above/below*, *left/right* is not at issue here (see e.g., Kemmerer, 2006).

An activation of these language-space assocations during word processing would provide evidence for experientially grounded semantic processing. This is exactly what has been found in many tasks (e.g., Bergen, Lindsay, Matlock, & Narayanan, 2007; Dudschig, Lachmair, de la Vega, De Filippis, & Kaup, 2012; Estes, Verges, & Barsalou, 2008; Gozli et al., 2013; Ostarek, Joosen, Ishag, de Nijs, & Huettig, 2019; Ostarek & Vigliocco, 2017; Richardson, Spivey, Barsalou, & McRae, 2003; Seger, Hauf, & Nieding, 2020; Verges & Duffy, 2009).

For example, participants had to react towards the color of a written word by pressing an upper or lower button. If the response direction matched the content of the presented words (e.g., the word *sun* written in blue font requiring an upward button press vs. the same word written in red font requiring a downward button press) participants responded faster (Lachmair, Dudschig, De Filippis, de la Vega, & Kaup, 2011). Presumably, when processing these nouns the actual (bodily) experiences during language acquisition are reactivated, like looking down to a frog or pointing upwards to a helicopter (Samuelson, Smith, Perry, & Spencer, 2011; Vogt, Kaup, & Dudschig, 2019). Thus, the spatial compatibility effects reported in the literature seem to be linked to simulations of situations in which objects typically occur. This view is supported by data from a study where participants had to discriminate between two target pictures (Ostarek & Vigliocco, 2017). A facilitatory effect was found when the target picture was presented in the vertical position on screen which was compatible to its typical location and when it was preceded by an up- or down-word belonging to the target's event. A spatially compatible control word not belonging to the target's event or a spatially neutral control word did not lead to facilitation. The authors conclude that rather abstract meaning features like UP or DOWN can not account for this finding. Moreover, spatial and oculomotor regions seem to be activated during the processing of implicitly spatial words, and the spatial properties of these words can be decoded by specific brain activity (Ostarek, Van Paridon, & Huettig, 2018).

Alternatively, some findings could also be explained by factors which are not directly related to sensorimotor experiences: The word pair *attic – basement* is processed faster when the words are vertically presented with *attic* above *basement* (Zwaan & Yaxley, 2003). However, this effect seems to be better explained by word order frequency instead of physically grounded language-space associations because the words *attic – basement* occur more often in this order than in the reverse order (Louwerse, 2008). Relatedly, it has been shown that not only direct personal experiences lead to spatial compatibility effects but that sensorimotor grounding of concepts can also happen indirectly when we learn novel concepts from language alone (Günther et al., 2020). For example, participants had to press one of two vertically aligned buttons to judge the sensibility of sentences with novel words like <*You scratch your mende*>. The newly learned referent for *mende* was either <enhanced head> or <bionic foot>. Depending on the spatial properties of the referent, spatial compatibility effects were observed. Moreover, the geographical location of cities in the real word can be reconstructed from distributional similarity values in texts because cities which are geographically closer tend to occur more often together in text (Gatti, Marelli, Vecchi, & Rinaldi, 2022; Louwerse & Zwaan, 2009). It remains to be shown whether the neural signatures of this indirect grounding are indeed equivalent to directly grounded semantic aspects (see Calvo-Merino, Grèzes, Glaser, Passingham, & Haggard, 2006, for neurally distinct signatures of observing actions we can perform ourselves versus actions which we only visually observe but can not perform).

Importantly, there are also some studies where researchers did not investigate whether the processing of words with mutual spatial content affects subsequent sensorimotor processes but where the order of testing was reversed – as it would necessarily be the case for language production paradigms. It was tested whether spatial cues facilitated an anagram task with anagrams from the semantic fields OCEAN and SKY, e.g., *dplhion* = *dolphin* or *cdulo* = *cloud*, (Berndt, Dudschig, & Kaup, 2018). Abstract spatial cues alone did not suffice to help solving the anagrams. However, providing more situational context with positioning the anagram in a spatially compatible location and presenting a background ocean-sky picture facilitated the task and participants needed less time to solve the anagrams.

Also using a visual stimulation, it was shown that vertical movement of dots on a screen interfered with a lexical decision task using motion verbs like *rise* or *fall* (Meteyard, Zokaei, Bahrami, & Vigliocco, 2008). Participants were slower to decide whether the presented words were real words or non-words when the semantic content of the words was incongruent to the movement direction, but only when the movement was near the threshold where it could be consciously perceived. Corroborating results for effects of visual motion on word processing have been reported by others (Dudschig, Souman, & Kaup, 2013; Kaschak et al., 2005). These findings are highly relevant for Study 1.

However, not only the direction of gaze and visual attention are important when investigating experiential traces of space in language. Body movements are also highly relevant because iconic gestures, pointing and body posture play a role during the acquisition of language-space associations (Öttl, Dudschig, & Kaup, 2017). Indeed, body orientiation seem to affect lexical access (Lachmair, Ruiz Fernández, et al., 2016). Participants were asked to recall a memorized word list of implicitly spatial words when they were placed on a rotating wheel and the wheel position changed so that their heads were either above or below their feet. They recalled significantly more often those words which were compatible with their own body position (see also Dijkstra, Kaschak, & Zwaan, 2007). This finding is highly relevant for Study 2.

All in all, there is a wide body of evidence that language-space associations are automatically activated when processing implicitly spatial language. These activations can be explained most economically by an embodied language processing account. Therefore, experiential traces of space are a suitable test case to study embodied language production.

1.2 Semantic processing during language production

1.2.1 Theoretical aspects

It is a core feature of spoken language production that speakers have a communicative intention and need to select the right words to get their message across. This idea is mirrored in many models of language production. They assume that a preverbal message needs to be translated into suitable lexical representations which in turn can be articulated (e.g., Abdel Rahman & Melinger, 2019; Caramazza, 1997; Dell, 1986; Jescheniak & Levelt, 1994; Levelt, 1989; Levelt, Roelofs, & Meyer, 1999; Roelofs, 2018).

The (preverbal) ideas are called concepts and they are organized at two levels according to the Featural and Unitary Semantic Space (FUSS) model (Vigliocco, Vinson, Lewis, & Garrett, 2004). The FUSS model contains many aspects to which language production researchers implicitly or explicitly subscribe. First, concepts consist of conceptual features which may be perceptual or sensorimotor but may also include taxonomic relations or functions of concepts.

The features are bound into unitary items, called lexical concepts. The links between features and concepts may be of different strength, depending on the relevance of the feature for the concept. This has been measured using feature listing tasks – features which have been listed by more participants have stronger links to the respective concepts (Vinson & Vigliocco, 2002). Also, conceptual features may be linked to several concepts at once and thereby establish similarity between concepts. Note that there are also some language production models which conceive of lexical concepts as unitary, non-decomposed items altogether (see Vinson et al., 2014, for a comparison of non-decompositional and compositional accounts of concepts in language production). However, in both accounts, there is a 1:1 mapping between a lexical concept and the semantically appropriate item from the mental lexicon, the so-called lemma. Lemmas already bear morpho-syntactical information like gender but are not specified phonologically. The links between features, lexical concepts and lemmas are bidirectional (e.g., Belke & Stielow, 2013). This may lead to co-activation of semantically related concepts and lemmas due to shared conceptual features, as activation spreads through the lexical-semantic network. In (picture) naming tasks, this may lead to co-activation of several potential candidates for verbalization. For example, when naming the picture of a <sparrow> a categorically related and visually similar concept like <chaffinch> might get co-activated due to feature-overlap. As a consequence of co-activation of lexical concepts, there is also co-activation at the lemma level. This may impede the selection of the target lemma *sparrow*.

Additionally, <sparrow> is also associated to concepts like <penguin> or <nest>. Whereas for example visual properties such as BROWN or SMALL are intrinsic features of <sparrow> which are shared with a closely related concept like <chaffinch>, concepts like <nest>, and <penguin> do not necessarily share perceptual features with the concept <sparrow>. Importantly, these concepts establish taxonomic, thematic or associative meaning relations (Kelter & Kaup, 2012) and they can also be co-activated in language production tasks.

Many theories assume lexical competition (Levelt et al., 1999), that is co-activated lemmas compete for selection. As a consequence, the activation of non-target lemmas influences how long it takes to select the target lemma, i.e. the lemmas for <sparrow>, <chaffinch>, <penguin> and <nest> compete for selection over time. Which lemma wins this competition and gets finally selected can be computed with mathematical rules (see e.g., Roelofs, 1992). Empirically, the impact of the co-activation of semantically related concepts and lemmas on language production has been investigated widely in picture naming tasks with voice onset latency as the main dependent measure. Usually, naming the picture of a <sparrow> in the context of a categorically related concept like <penguin> slows naming in comparison to naming the picture of a <sparrow> in the context of an unrelated concept like <chair>. This relative slowing has been called interference. In contrast, naming the picture of a <sparrow> in the context of a thematically related concept like <nest> usually speeds up naming in comparison to naming a <sparrow> in the context of an unrelated concept like <knife>, which has been called facilitation (see below for empirical evidence). Thus, differences in the timing of picture naming depending on context manipulations have been the main focus of investigating semantic processing in language production (see Section 1.2.2). It has also been suggested that conceptual activation and lexical activation are interactive processes during language production (e.g., Abdel Rahman & Melinger, 2019), with evidence showing that conceptual activation starts earlier than lexical activation (Carota, Schoffelen, Oostenveld, & Indefrey, 2022; Indefrey, 2011; Strijkers & Costa, 2016).

After lemma selection, the word form (lexeme) is passed on to the next stages of language production: the selected word form gets morphologically specified and phonologically encoded before being articulated (e.g., Levelt et al., 1999).

1.2.2 Empirical investigations of semantic relations in language production

Selecting the right lemma from among a cohort of co-activated lexical entries has been investigated using different context paradigms. In the picture word interference task (PWI), to-be-named pictures are presented together with distractor words (e.g., Bürki, Elbuy, Madec,

& Vasishth, 2020; Glaser & Düngelhoff, 1984; Lupker, 1979). In the semantic blocking paradigm, pictures are presented in blocks with semantically homogeneous or heterogeneous pictures (e.g., Belke, Meyer, & Damian, 2005; Damian, Vigliocco, & Levelt, 2001; Kroll & Stewart, 1994) and in the continuous naming task they are presented in a row with (seemingly unrelated) other pictures to investigate sequential effects of naming (e.g., Howard, Nickels, Coltheart, & Cole-Virtue, 2006). The semantic relationship between the target and the naming context is varied, with category membership having most often been used as an operationalisation of semantic relatedness.

Across paradigms, it was found that naming pictures in contexts with categorically related concepts leads to slower naming latencies compared to naming the same pictures in contexts with unrelated concepts (e.g., Costa, Strijkers, Martin, & Thierry, 2009; Damian et al., 2001; Roelofs, Piai, & Schriefers, 2013). This effect is called semantic interference. While many researchers agree that the co-activation of related lexical concepts leads to competition and thus to the observed semantic interference effect, other mechanisms have been proposed, too (e.g., Abdel Rahman & Melinger, 2019; Belke & Stielow, 2013; Oppenheim, Dell, & Schwartz, 2010; Roelofs, 2018; but see Mahon, Costa, Peterson, Vargas, & Caramazza, 2007). The concrete mechanism is not important with regard to this dissertation and will therefore not be discussed further. However, it should be noted that most researchers agree that the observed effect of slowed naming is indicative of lexical access processes.

Importantly, semantic interference effects were also found when participants did not name the presented pictures but classified the last letters of the target words as a vowel or consonant via button press (Abdel Rahman & Aristei, 2010; Hutson, Damian, & Spalek, 2013; Tufft & Richardson, 2020). The effect was of similar size as the typical PWI effect, which is considerably small (about 20 ms) but reliably observed according to a recent metastudy (Bürki et al., 2020). Thus, not only overt naming but also tasks which require access to lemmas and their orthographical form can be used to study lexical access. This is an important insight we used for Study 3.

In many studies, categorical relationships were manipulated dichotomically with targets either being categorically related to the naming context or not (e.g., Abdel Rahman & Aristei, 2010; Glaser & Düngelhoff, 1984; Howard et al., 2006; Hughes & Schnur, 2017; Kroll & Stewart, 1994; Piai, Roelofs, & Van Der Meij, 2012). However, some did not manipulate categorical relatedness dichotomically and it has been observed that an increase in the amount of semantic similarity (operationalized as shared semantic features between a target picture and categorically related distractors) leads to more interference (Rose & Abdel Rahman, 2017; Rose, Aristei, Melinger, & Abdel Rahman, 2019; Vieth, McMahon, & de Zubicaray, 2014; Vigliocco, Vinson, Damian, & Levelt, 2002).

Importantly, semantic relatedness is not confined to categorical relations. Yet, fewer studies investigated semantic relations apart from category membership (for an overview see Abdel Rahman & Melinger, 2019). Investigating thematic or associative links, semantic context effects were observed, too. There, facilitation as well interference have been observed depending partly on the specific manipulations used (see e.g., Abdel Rahman & Melinger, 2007; Aristei, Melinger, & Abdel Rahman, 2011; Damian & Spalek, 2014; de Zubicaray, Hansen, & McMahon, 2013; La Heij, Dirkx, & Kramer, 1990). Interestingly, naming seemingly unrelated words like *dandelion*, *shoe* and *cereal bar* induced interference in a blocking paradigm when the blocks were preceded by a title or a story which provided a common theme like HIKING TRIP (Abdel Rahman & Melinger, 2011; Lin, Kuhlen, & Abdel Rahman, 2021). In the continuous naming paradigm, interference was even observed for thematically related objects without providing information about their context (Lin, Kuhlen, & Abdel Rahman, 2022; Rose & Abdel Rahman, 2016). This demonstrates that participants use contextual relations flexibly and quickly infer even shallow semantic relations or build ad-hoc categories (Barsalou, 1983).

Lately, researchers started to move away from bare picture naming tasks. For example, they manipulated the social character of the naming context (e.g., Gambi, Van de Cavey, & Pickering, 2015; Kuhlen & Abdel Rahman, 2017; Lin et al., 2021). Furthermore, they provided richer semantic context. Participants were asked to name pictures after reading semantically constraining versus non-constraining sentences while the continuous EEG was recorded (Hustá, Zheng, Papoutsi, & Piai, 2021). Naming after constraining sentences led to facilitation as well as alpha-beta desynchronizations in the EEG. This was interpreted as a signature of target word prediction. This paradigm also closely combined sentence comprehension and word production which is a typical feature of everday language use (Pickering & Garrod, 2013). Therefore, this paradigm is a move forward to studying language production in a more ecologically valid way and it is similar to the task used for Study 1 and 2.

Critically, the semantic aspects which have been looked at are only a small part of the manyfold meaning aspects which might play a role when formulating verbal messages. Researchers so far focused mainly on categorical relations and to a lesser degree on associative-thematic relations. However, emotional and social meaning aspects as well as meaning aspects based on sensorimotor experiences have received little consideration. The following section provides an overview of the empirical work investigating the role of embodied meaning aspects in language production.

1.2.3 Evidence for embodiment in language production

Evidence for a role of embodied meaning aspects in language production is scarce. Yet it has been shown that experiences based on different sensory modalities may get reactivated. For the visual domain it has been demonstrated that naming greyscale pictures of high-color diagnostic objects like <banana> is slower when preceded by a matching color prime in comparison to naming low-color diagnostic objects like <bicycle> (Redmann, FitzPatrick, Hellwig, & Indefrey, 2014). Moreover, shared visual shapes irrespective of other shared semantic relations lead to interference in the PWI and a blocking task (de Zubicaray et al., 2018). For example, a black-and-white drawing of the target word *igloo* was paired with one of two semantically unrelated distractor words like *turtle* and *feather.* When the visual shape of the distractor word's referent resembled the visual shape of the target word's referent (e.g., the combination *igloo – turtle*) participants showed slower naming latencies compared to naming the target with the other distractor word not sharing visual features (e.g., the combination *igloo – feather*).

For the auditory domain, it has been shown that naming of objects, which are highly associated with a sound like <dog> or <telephone>, slowed down compared to naming objects not associated with sounds when the pictures were presented together with white-noise bursts. The white-noise presumably interfered with accessing the sound features which are important to process the respective concepts (Mulatti, Treccani, & Job, 2014). It was also shown that sounds which are associated with concepts can influence their activation in a variation of the PWI paradigm (Mädebach, Wöhner, Kieseler, & Jescheniak, 2017). Here, target pictures were presented together with distractor sounds. Presenting animal pictures with a coherent sound (e.g., <horse> + a neighing sound) led to facilitation whereas sounds associated with semantically related concepts (e.g., <horse> + a barking sound) led to interference relative to a control condition with unrelated sounds (e.g., <horse> + a drumming sound). Relatedly, research with aphasic patients showed that multisensory cues lead to facilitatory effect. The patients received semantic auditory cues which were related to the targets they had to name (e.g., a ringing sound before naming <telephone>). This intervention improved naming accuarcy and can be accounted for by sensorimotor theories of language processing according to the authors (Grechuta et al., 2020).

As for language comprehension, most of the work on embodied meaning aspects in language production investigated the motor content of concepts. For example, the words *close* and *open* were presented to participants while they were engaged in a concurrent motor task which consisted of either opening or closing their hands following a color cue (Liepelt, Dolk, & Prinz, 2012). Participants also had to produce the words *close* or *open* after a color cue while simultaneously viewing a picture of an opening or closing hand movement. When there was a semantic mismatch between the word and the movement they found interference, both when participants produced the movement and when they produced the respective words, suggesting that there is a bidirectional functional link between action and language.

Interestingly, naming words with a high degree of association to particular body movements like *shovel* showed no decline in naming accuracy with age whereas for naming words with no association to body movements like *panda* accuracy declined with higher age (Reifegerste, Meyer, Zwitserlood, & Ullman, 2021). This finding supports the idea that experiential traces are linked to semantic processing in language production and shows that sensorimotor experiences may alleviate age related decline in language processing.

Neurally, it has been shown that repetitive transcranial magnetic stimulation (rTMS) pulses over the left inferior parietal lobe (IPL) of healthy participants leads to slower naming of highly manipulable objects compared to naming non-manipulable objects or living entitites (Pobric, Jefferies, & Lambon Ralph, 2010a). It was concluded that the suppression of the sensorimotor information (here the praxis information coded in the IPL) affects the processing of concepts for which these information are central. In a similar logic, participants were asked to name depicted tools while performing a secondary task, namely squeezing a ball with their hand (Witt, Kemmerer, Linkenauger, & Culham, 2010). Slower naming was found when the handles of the depicted tools faced the squeezing hand (presumably because it was already occupied) in comparison to naming animal pictures. However, this finding was not replicated by a different group of researchers who found that squeezing a sponge hindered access to both tool and animals names (Matheson, White, & McMullen, 2014). Yet there is more evidence suggesting that an occupation of the motor system interferes with naming target words with motor content: In an object naming task, participants performed manual movements which would have made interaction with the depicted objects impossible. More naming errors were found the more experience participants had in manipulating the depicted objects (Yee, Chrysikou, Hoffman, & Thompson-Schill, 2013). Thus, not only the motor content but also the individual experiences of participants influenced the strength of the embodiment effect. Interference has also been found when participants had to produce verbs for depicted hand- or foot-related actions while they were engaged in a motor task with either their hands or their feet (Hirschfeld & Zwitserlood, 2012). When the effector of the verb action matched the effector used for the motor task, slower naming was observed. In a second experiment they also showed interference using a classic language production paradigm by blocking action pictures according to their effectors in either homogeneous (hand vs. feet only actions) or heterogeneous blocks (see also de Zubicaray, Fraser, Ramajoo, & McMahon, 2017, for evidence in a similar setup using fMRI). Furthermore, populations with selective motor deficits (like Parkinson's disease) exhibit problems in producing verbs for depicted actions when they have a high motor content (Herrera, Rodríguez-Ferreiro, & Cuetos, 2012, see also Klostermann et al., 2022).

Importantly, not only interfering but also facilitating effects of motor actions have been found. While sequentially pressing five buttons with their fingers from thumb to pinky, participants had to produce number words for numbers which were presented on a screen (Sixtus, Lindemann, & Fischer, 2018). Shorter production latencies were found when the finger pressing the button matched the finger which represents the to-be-named number when counting with fingers. Moreover, in random number generation tasks participants reliably produce larger numbers when their bodies or their eyes move upwards in a vertical direction (Hartmann, Grabherr, & Mast, 2012; B. Winter & Matlock, 2013). Thus, the bodily foundations of mathematical processing are accessed when producing number words.

Interestingly, not only experiential meaning aspects directly related to our body and to our senses are reflected in semantic processing during language processing. It has also been found that the speed of objects influences how fast we name pictures of these (Ben-Haim, Chajut, Hassin, & Algom, 2015) with speed being an experientially grounded feature. The naming latencies for words contained in databases as the International Picture Naming Project (Szekely et al., 2004) and the English Lexicon Project (Balota et al., 2007) were predicted by the rated speed of objects over and above lexical factors which were controlled for (e.g., frequency, name agreement, word length). Furthermore, movement speed of objects was experimentally manipulated. For example, a <car> was either depicted moving uphill or downhill. Naming latencies were slower when the car was moving uphill and therefore supposedly moving slower compared to when it was moving downhill and supposedly fast.

Taken together, the few existing studies suggest that signatures of processing embodied meaning aspects in language production exist. However, some of the above studies are not explicitly framed in that regard. In other words, semantic aspects related to sensorimotor experiences are manipulated and naming responses are analysed but the findings are not discussed in relation to experientially grounded meaning processing (e.g., de Zubicaray et al., 2018; Mädebach et al., 2017; Witt et al., 2010) or they are not related to theories of language production (e.g., Sixtus et al., 2018; Witt et al., 2010). Therefore, it remains unclear what kind of semantic architecture could be causing the effects. While it could be argued that reactivations of experiential traces can account for the above findings, there are only some attempts to relate the findings to both embodied cognition and semantic theories of language production (Ben-Haim et al., 2015; de Zubicaray et al., 2017; Hirschfeld & Zwitserlood, 2012; Mulatti et al., 2014). In one of the few studies where this is explicitly done, the authors interpret their findings cautiously: Hirschfeld and Zwitserlood (2012) argue that their findings might not only be explained by a theory of embodied meaning processing but might also be accounted for by amodal and abstract semantic features. For example, abstract semantic features like *has toes*, *can be moved*, *used for walking* might get activated when we name actions executed with the feet and may lead to competition between co-activated lexical nodes sharing these features.

Moreover, few studies which have been presented in this section employed classic language production paradigms such as the PWI, blocking tasks or continuous naming. This makes it difficult to link findings of embodied meaning processing in language production to semantic context effects found for categorical or thematical relations (see Section 1.2.2).

Note that most of the studies investigating semantic processing used naming tasks where the timing of auditory responses served as dependent variable. Therefore, the availability of reliable implementations of timing-sensitive picture naming paradigms is of utmost importance for investigating semantic processing in language production. With the following section I show how this prerequisite posed a sincere challenge for this dissertation in times of Corona when testing under controlled conditions in the lab was not possible.

1.2.4 Investigating language production in times of Covid-19

During the Covid-19 pandemic, many labs all over the world had to be closed. As lab-based testing was thus no longer possible, we were eager to move our paradigms online. Running psychological experiments online brings advantages compared to testing participants in the lab. For example, it is easier to access bigger and more diverse samples (Grootswagers, 2020). However, language production studies relying on analysis of speech data had not been conducted online prior to the beginning of the Covid-19 pandemic (see also Fairs & Strijkers, 2021).

The main concern with running online language production experiments relying on overt articulation is the so-called audio-visual synchrony problem: It has been observed that the reliability of presenting visual and auditory stimuli at the same time in web-based experiments is poor (Bridges, Pitiot, MacAskill, & Peirce, 2020; Reimers & Stewart, 2016). To date, the basis technology for online experiments (Java Script) does not allow for a precise timelocking of presenting pictorial stimuli and recording audio responses (Bridges et al., 2020). Therefore, in a web-based experiment, it is difficult to fully control for stimulus-locked onsets of audio recordings. This poses a problem for studies investigating semantic processing in language production where usually pictures are presented and subsequent auditorily responses are analyzed with the latency of spoken responses serving as main dependent variable. Thus, there is a need for reliable implementations of timing-sensitive online picture naming paradigms, which I addressed with Study 3.

1.3 Open questions and outline of the present work

With regard to this dissertation, three aspects are notable when zooming in on semantic relations in language production. First, apart from the few exceptions mentioned in Section 1.2.3, literature on the role of different meaning aspects in language production is scarce. The main focus was on categorical and to a lesser degree thematically relations. Second, the specific time course of conceptual and lexical activation has been investigated (Indefrey, 2011; Strijkers & Costa, 2016) and the debate has focused on the specific mechanism which may explain interference versus facilitation (e.g., Abdel Rahman & Melinger, 2009; Mahon et al., 2007; Oppenheim et al., 2010). At the same time, current models of lexical-semantic encoding in language production do not reflect insights from neuropsychology about the organization of semantic memory (Belke, 2013). Neither embodied meaning aspects nor distributional accounts of semantics are integrated in current models and an emcompassing account of the architecture of the semantic system used for language production is missing (see e.g., Vinson et al., 2014, for a call to integrate insights from current theories of semantics in language production research). For example, researchers refer to a semantic effect as being "associated with conceptual and lexical retrieval" (Hustá et al., 2021, p. 10) without differentiating between the two. Third, the investigations of language production have mainly focused on differences in the time course of producing predefined target words. However, to find out more about the human capacity to speak and not only about the preconditions of obtaining different voice onset latencies, we need to focus on the core feature of language production again. This is our ability to choose a suitable word to express an intended meaning. Therefore, it would be necessary to find out which words are actually chosen (and which words not). To do exactly that, the paradigm for Study 1 and 2 was developed.

- With Study 1 and 2 the following research question was addressed: Is there evidence for experientially grounded meaning aspects influencing lexical choices during language production? Furthermore, it was explored in how far semantic relatedness measures based on linguistic distributional measures of semantics interact with experientially grounded meaning aspects. In both studies the role of experiential traces of space was investigated. So far they had not been investigated in language production. Study 1 places the focus on the reactivation of these aspects through the visual modality. In Study 2 bodily movements are manipulated by using body posture changes (head up vs. head down). Thus, Study 2 follows up on Study 1 by introducing a manipulation which should lead to a stronger reactivation of experiential traces compared to a mere visual stimulation. Hitherto, the focus in empirical investigations of semantic processing in language production has been laid on the time course of producing predefined words. Importantly, Study 1 and Study 2 use a novel paradigm which allows to empirically investigate with quantitative measures WHAT people are saying in a relative unconstrained language production task.

- With Study 2 we also aimed to answer the research question whether interindividual differences in the susceptibility towards the bodily manipulation may influence lexical selection. To this end we explored in how far interindividual differences in interoceptive sensibility influence semantic processing.

- With Study 3, we show how classic language production experiments, which rely on timed spoken responses, can be conducted when access to participants is difficult and we demonstrate how lab-based language production experiments can be transferred to the internet. The aim was to provide a proof of concept for running online language production experiments by (1) demonstrating that well established language production effects can be replicated online, and by (2) comparing a vocal response to a manual button press response which can be more easily implemented in an online experiment. We also aimed to show that even experiments requiring a high degree of control over participants can be conducted online. To this end, in Study 2, which was conducted after Study 3, the implementation for online audio recording was combined with a parallel video call session. This allowed the experimenter to control the execution of the body posture changes.

In Chapter 2 of this dissertation a summary of the three empirical studies is presented. Study 1 has been accepted and published online by the *Quarterly Journal of Experimental Psychology* (Vogt, Kaup, & Abdel Rahman, 2022), Study 2 has been submitted for publication and is currently under review. Study 3 has been published by the journal *Behavior Research Methods* (Vogt, Hauber, Kuhlen, & Abdel Rahman, 2022). All studies except for Experiment 1 in Study 1 have been preregistered. Material, data and analysis scripts are made available online for all the studies. An author's version of each manuscript can be found in Chapter 3 before the findings are jointly discussed in Chapter 4. Based on the insights from this dissertation, I also discuss the need for better theories of semantics in language production and provide ideas for what they should entail.

2 Summary of the present studies

2.1 Study 1: Experience-driven meaning affects lexical choices

With this study, we investigated whether reactivating experiential traces of space has an effect on lexical choices of participants in a sentence completion task. The experiential reactivation of vertical space was implemented by presenting sentences word by word on a screen in rapid serial visual presentation mode. The important manipulation was that each word appeared slightly below the previous word (or slightly above respectively) thereby inducing an upward or downward movement of sentences on screen. Participants had to complete the sentences with a suitable noun phrase and their responses were recorded.

We expected that the visual movement of the sentences would reactivate experiential traces of space, and would lead to words being selected in accordance with the spatial movement of the sentences. Thus, we expected that words would refer to objects higher up in space after upward compared to after downward moving sentences. The sentence fragments described situations across a wide range of vertical spatial locations, like <*You are walking along the beach and you see a...*> or <*You are hiking in the mountains and you see a ...*>. We also explored whether the spatial locations conveyed by the stimulus sentences influenced lexical choices. Moreover, we wanted to find out if and in how far the spatial properties of the produced nouns would depend on an interaction between the spatial properties of the presented sentences and a certain level of semantic similarity between the produced nouns and the nouns in the sentences. To this end, the semantic similarity between the produced nouns and the nouns in the stimulus sentences was computed using a distributional measure of semantic similarity (Günther, Dudschig, & Kaup, 2015).

We ran the experiment with native German speakers ($n_{\text{Exp1}} = 33$). Subsequentely, we conducted the experiment a second time with small adjustments to replicate the results found in Experiment 1 ($n_{\text{Exp2}} = 72$). We used 90 sentence fragments in Experiment 1 and 60 sentence fragments in Experiment 2, with sentences spanning a wide range of vertical locations from down in space (e.g., <*You are standing at a lake and you see...*>) as well as higher up in space (e.g., <*You are hiking in the mountains and you see...*>). All sentences were designed in such a way that participants could complete them with a great variety of words. In Experiment 2, we also introduced a baseline condition where sentences were displayed word by word in the center of the screen. Furthermore, the sentence fragments were presented without a determiner before the noun phrase to make the task easier for participants (i.e. in Experiment 1 participants read the sentence <*You are standing at a lake and you see* ***a***...>, in Experiment 2 they read the sentence <*You are standing at a lake and you see ...*>). After the word production phase, a different set of participants ($n_{\text{Exp1}} = 35$, $n_{\text{Exp2}} = 30$) rated the produced nouns according to the spatial location of their referents on a 7-point Likert scale (1 = low, 7 = high).

Contrary to our expectations, the vertical visual movement of sentences on screen did not influence lexical choices of our participants. However, they chose nouns as suitable sentence completions whose spatial properties were influenced by the spatial locations described in the sentences. For example, participants were more likely to produce a word like *shell* compared to a word like *gull* after a sentence like <*You are walking along the beach and see...*> as both *beach* and *shell* are located down in space. Importantly, an answer like *gull* would have been semantically as similar to *beach* as *shell* according to a distributional measure of semantic similarity (Günther et al., 2015), but was chosen less. Importantly, the spatial properties of the produced nouns were predicted by the spatial properties of the presented sentences across all levels of semantic similarity between the presented sentences and the produced words, albeit to different degrees. Thus, even in cases of low semantic similarity between the produced word and the noun in the stimulus sentence, namely when participants produced words which were not highly associated with the displayed sentences, there was an effect of the spatial properties of the presented sentences on the spatial properties of the produced words.

Taken together, we interpret these findings as evidence for a reactivation of the experientially grounded meaning dimension of vertical space. More importantly, these results show that the reactived experientially grounded meaning affects which lexical candidate is selected. Furthermore, the finding that semantic similarity and the reactivation of experiential traces of space which are embedded in text both contribute to the observed outcome is in line with recent work on semantics. Accordingly, experientially grounded meaning aspects and linguistic distributional aspects of meaning are highly intertwined yet distinct aspects of meaning (see e.g., Banks, Wingfield, & Connell, 2021; Carota, Nili, Pulvermüller, & Kriegeskorte, 2021; Davis & Yee, 2021).

However, there was no effect of vertical visual movement of the sentences on lexical choices of participants. Potentially, the visual movement was too abstract and thus did not suffice as a semantic cue during the lexical selection process narrowing down the potential set of lexical candidates. Previous studies also showed that spatial cues needed to encompass a high degree of context in order to influence lexical processing in an anagram solving task (Berndt et al., 2018) and that they need to be sufficiently salient to elicit reactivations of language space associations (Dudschig & Kaup, 2017).

In a nutshell, Study 1 showed that experiential traces of space as incorporated in linguistic constructions may influence lexical choices. Our results thus provide tentative evidence for experientially grounded lexical selection.

2.2 Study 2: Sensorimotor activations and interoceptive sensibility influence which words we choose when speaking

Building on the paradigm and the findings from Study 1, we aimed to further investigate the reactivation of experiential traces of space and their influence on lexical choices in language production. We again asked our participants to complete sentence fragments. Presumably, the visually induced movement of the presented sentences in Study 1 did not influence lexical choices because the manipulation was too abstract. Therefore, we aimed to increase the sensorimotor activation of participants in Study 2. Here, sentence fragments were presented auditorily and participants had to perform an upward or downward head movement before completing them.

We expected that the head movements would reactivate experiential traces of space. This, in turn, should lead to participants producing words that share the spatial properties of the head movement they just performed. That is, the spatial locations of the words' referents should be influenced by the head movement direction.

Interoceptive sensibility, namely the general tendency to perceive body processes (Mehling et al., 2012), seems a likely candidate for interindividual differences in a paradigm manipulating body posture and investigating the impact of body movement on lexical choices. Therefore, two facets of interoceptive sensibility were assessed: participants' ability to sustain and control attention to body sensations (*attention regulation*) and the awareness of uncomfortable, comfortable and neutral body sensations (*noticing*). We expected to find a moderating influence of interoceptive sensibility on potential effects of head movement direction on the choice of suitable nouns.

Additionally, we expected to replicate the findings from Study 1, namely that the spatial properties of produced nouns can be predicted from the spatial locations conveyed by the stimulus sentences.

We conducted the experiment with native German speakers ($\mathrm{n}_{\mathrm{Exp1}} = 44$) and replicated it with small procedural changes ($\mathrm{n}_{\mathrm{Exp1}} = 55$). We used 42 auditorily recorded sentence fragments, which had already been used in Study 1. The experiment itself was run online with participants entering a video call session with the experimenter who then sent a link for participation. The experimenter stayed in the video call throughout the word production phase to instruct participants and to observe the correct execution of the head movements. Trials started with an auditory cue indicating the head movement direction (up vs. down). Participants then closed their eyes and moved their heads upwards or downwards while the sentence fragments were auditorily presented. Participants kept their head in the head-up or head-down position when the fragments ended and produced a noun as suitable ending. Any noun which was produced in a time frame of 5 sec after the sentence fragment ended was recorded and considered for later analysis. Participants also filled out two subscales from the Multidimensional Assessment of Interoceptive Awareness, *attention regulation* and *noticing* respectively (Mehling, Acree, Stewart, Silas, & Jones, 2018; Mehling et al., 2012). Afterwards, a different sample of participants ($n_{\text{Exp1}} = 31$, $n_{\text{Exp2}} = 32$) rated the produced words according to their spatial properties on a 7-point Likert scale (1 = low, 7 = high).

The procedure between Experiment 1 and 2 differed in four ways: First, the tone-response assignment was reversed. A high tone indicated upward head movement in Experiment 1, and downward head movement in Experiment 2. Second, participants received the interoception questionnaire after the word production phase in Experiment 1 and before the word production phase in Experiment 2. Third, in Experiment 2 all produced words were not only rated by an independent group of raters but each participant also rated the words they had produced themselves. Finally, because mindfulness has been related to similar aspects as interoception (Gibson, 2019), participants also received a mindfulness questionnaire in Experiment 2 (FFMQ: Baer et al., 2008).

In line with our expectations, the head movements influenced the lexical choices. Upward head movements led to nouns which were located higher up in space than nouns produced after downward head movements and vice verse. This effect failed to reach significance in Experiment 1, but was significant in Experiment 2, although the effect was numerically similar in both experiments. Furthermore, we found an interaction between the head movement direction and participants' ability to sustain and control attention to body sensations (*attention regulation*) in Experiment 1, which failed to reach significance in Experiment 2. No effect was found for the awareness of uncomfortable, comfortable and neutral body sensations (*noticing*). The differences between Experiment 1 and 2 can presumably be attributed to the reversed order of administering the word production task and the interoception questionnaires. Starting with the interoception questionnaire in Experiment 2 may have raised awareness to bodily signals in general, thereby levelling out the impact of interindividual differences. However, pooling the data of Experiment 1 and 2 stabilized the effects and both the main effect for head movement direction as well as the interaction between *attention regulation* and movement were significant across both experiments.

We also replicated the findings from Experiment 1 and showed that the produced nouns bear the spatial characteristics of the wider sentence context. There was an interaction between semantic similarity and spatial locations of nouns in the stimulus sentences. With higher semantic similarity the effect of spatial locations of sentence nouns on the spatial properties of the produced nouns was enhanced. However, even for loosely related nouns the spatial properties of the sentence context predicted the spatial properties of the produced nouns.

In Experiment 2, the scales of the mindfulness questionnaire did not interact with the head movement manipulation and were therefore not considered further. Instead we focused on those facets which had been collected in both Experiment 1 and Experiemnt 2 (*attention regulation* and *noticing*). Furthermore, the exploratory analysis of the spatial rating done by participants who produced the words themselves did not show a significant impact of the head movements on the spatial properties of the produced nouns. However, the effect went in a similar direction as the effect in the main analysis.

Taken together, the findings show that the experientially grounded meaning dimension of vertical space can be reactivated via bodily movements. These then affect the lexical choices we make. We complement this result by showing that interinvidivual differences in a measure of interoceptive sensibility may moderate embodied language processing. Participants with higher scores in *attention regulation* were more susceptible to the head movement manipulation.

2.3 Study 3: Internet-based language production research with overt articulation

With this study we demonstrate the feasability of conducting online picture naming experiments using voice onset latencies from audio recordings as dependent variable. It is the first study to show that voice onsets can reliably be used when collected online in within-subject designs. Raw data as well as analysis scripts and code for an example experiment based on jspsych (de Leeuw, 2015) are freely available over the Open Science Framework.

We used the picture word interference (PWI) paradigm. In this paradigm, participants have to name pictures with visually superimposed distractor words which have to be ignored. Typically, naming latencies get longer when there is a (categorical) semantic relation between the target and the distractor word. For example, naming of the word *dog* is slower when the distractor word is *cat* compared to a distractor like *cup* (e.g., Lupker, 1979; Schriefers, Meyer, & Levelt, 1990). Importantly, similar semantic interference effects are found when participants classify the last letter of the target word. That is, they have to press a keyboard button to decide whether the last letter represents a vowel or a consonant (Abdel Rahman & Aristei, 2010; Hutson et al., 2013; Tufft & Richardson, 2020).

In a series of three online experiments with 48 participants each, we conceptually replicated the study by Abdel Rahman and Aristei (2010). Participants took part in both a naming task and a vowel consonant classification task in a PWI paradigm. For the overt naming task, we customized code for recording audio on the web to implement trials where participants saw the picture and time-locked audio recordings were started for each trial. Note that due to the audiovisual synchrony problem, the experimenter can not control whether the audio recording starts at exactly the same moment when the picture is presented. Afterwards, the recordings were preprocessed and voice onset timings were calculated using Chronset (Roux, Armstrong, & Carreiras, 2017) and Praat (Boersma & Weenink, 2020).

We found a semantic interference effect both in the vowel consonant classification task as well as in the overt naming task across the three experiments. Importantly, we found an effect size of around 20 ms which is similar to those published in the literature (for a meta-analysis see Bürki et al., 2020). Thus, we demonstrated that this well established semantic effect is comparable in online and lab-based experiments. With the vowel consonant classification task, we validated the semantic interference effect in overt naming against a measure which used button press latencies. Note that the reliability of response latencies from keyboard button presses has already been proven in online settings (see e.g., Hilbig, 2016; Pinet et al., 2017). Replicating an established effect relying on overt responses and validating its timing with button presses demonstrates that our implementation of recording time-locked audio provides a suitable and feasible way to conduct language production experiments with overt responses online.

A post-hoc power analysis complemented the picture by providing estimates for sample size and trial number. These can be used for future projects from researchers wanting to move their language production paradigms online. Moreover, they are helpful to further underline the validity of data collected online and to check how many resources planned online language production experiments might need. This is especially important because online language production experiments relying on overt naming responses do not yet provide a more efficient way of data collection due to the time-consuming audio data preprocessing before analysis.

As our data show that good and reliable results can be achieved in online picture naming studies, our implementation of recording audio can be considered a suitable research tool. Based on our experience, we also provide suggestions for running online (language production) experiments which will raise the data quality while minimizing the resources for online testing on the side of the researchers (see also Section 4.4).

3 Original research

3.1 Manuscript 1: Experience-driven meaning affects lexical choices during language production

This manuscript was published as:

Vogt, A., Kaup, B., & Abdel Rahman, R. (2022). Experience-driven meaning affects lexical choices during language production. *Quarterly Journal of Experimental Psychology.* Advance online publication. https://doi.org/10.1177/17470218221125425

This is the final author copy with slight formatting adjustments for inclusion in this dissertation.

Appendix A and B can be found in an online repository, together with data and analysis scripts:

https://osf.io/se6a3/?view_only=531a16b927a54b558d340f526148b881

References are listed at the end of the dissertation.

Abstract

The role of meaning facets based on sensorimotor experiences is well-investigated in comprehension but has received little attention in language production research. In two experiments, we investigated whether experiential traces of space influenced lexical choices when participants completed visually-presented sentence fragments (e.g., <*You are at the sea and you see a ...*>) with spoken nouns (e.g., *dolphin, palm tree*). The words were presented consecutively in an ascending or descending direction, starting from the center of the screen. These physical spatial cues did not influence lexical choices. However, the produced nouns met the spatial characteristics of the broader sentence contexts such that the typical spatial locations of the produced noun referents were predicted by the location of the situations described by the sentence fragments (i. e., upper or lower sphere). By including distributional semantic similarity measures derived from computing cosine values between sentence nouns and produced nouns using a web-based text corpus, we show that the meaning dimension of LOCATION IN SPACE guides lexical selection during speaking. We discuss the relation of this spatial meaning dimension to accounts of experientially grounded and usage-based theories of language processing and their combination in hybrid approaches. In doing so, we contribute to a more comprehensive understanding of the many facets of meaning processing during language production and their impact on the words we select to express verbal messages.

Keywords: Language Production, Experiential Traces, Language Grounding, Hybrid Models, Lexical Selection, Semantic Processing

A central process during language production is the selection of the right words to express an intended meaning. While the role of some meaning aspects – like categorical relations – is well investigated, little is known about others (Abdel Rahman & Melinger, 2019). Specifically, and in contrast to language comprehension, little is known about meaning aspects grounded in sensorimotor experiences. This is surprising because we frequently talk about our sensations and experiences in everyday life. Therefore, meaning aspects linked to our sensory experiences seem fundamental in language production.

The present study was designed to investigate influences of experientially grounded meaning on lexical-semantic processing during language production. Furthermore, we relate sensorimotor experiences to a measure of semantic similarity by using linguistic distributional measures of meaning relations.

Semantic relations in language production

When speakers plan to produce a message, meaning representations at the conceptual level and word representations at the lexical level (lemmas) – as well as semantically related conceptual and lexical entries – are activated, and the target lemma is selected from among these co-activated alternatives (Caramazza, 1997; Dell, 1986; Levelt et al., 1999; Mahon et al., 2007; Oppenheim et al., 2010). Evidence of lexical-semantic factors influencing lexical selection stems from context effects induced by displaying constraining vs. non-constraining sentences before asking individuals to name a picture (Hustá et al., 2021), from context effects by previously named related pictures (e.g., in the cyclic blocking and continuous naming paradigm; Belke et al., 2005; Howard et al., 2006), or simultaneously-presented related distractor words (in the picture word interference paradigm; e.g., Glaser & Düngelhoff, 1984; for a recent discussion see Bürki et al., 2020; Roelofs, 2018). Typically, categorical semantic relations have been investigated in these paradigms. However, the meaning of verbal messages is multifaceted and may as well contain information about associations, part-whole-relations, thematic links, as well as social and emotional meaning aspects. Therefore, it should not be reduced to categorical relations (Abdel Rahman & Melinger, 2019; Jackson, Hoffman, Pobric, & Lambon Ralph, 2015), but investigations of non-categorical relations during lexical selection are comparatively rare and have focused on thematic, situational or associative relations (Abdel Rahman & Melinger, 2009, 2019; Alario, Segui, & Ferrand, 2000; Aristei & Abdel Rahman, 2013; Costa, Alario, & Caramazza, 2005; Damian & Spalek, 2014; de Zubicaray et al., 2013; La Heij et al., 1990; Lin et al., 2021; Rose & Abdel Rahman, 2016). Crucially, lexical-semantic processing is not confined to traditionally investigated semantic relations, and may include a much wider range of meaning facets based on sensory experiences like aspects of sound, shape and color which have been shown to play a role during language production (de Zubicaray et al., 2018; Mädebach et al., 2017; Redmann et al., 2014).

Experientially grounded representations in language comprehension

Experiential grounding refers to the idea that the multimodal – and often bodily – experiences we have made leave experiential traces in our brain and become tied to our knowledge about these objects, situations or actions (Barsalou, 2008) and, consequently, to the linguistic constructions and words used in those situations (e.g., Lynott et al., 2020; Zwaan & Madden, 2005). Due to its strong link to bodily sensations, this line of work is often referred to as embodiment or embodied cognition. We use the phrase 'experiential grounding' throughout this paper in order to highlight that not all experiences are based on bodily sensations. From this perspective, concepts can be understood as modality-specific, experience-dependent and flexible representations in distributed neural networks which include, but are not restricted to, sensorimotor areas of the brain (Kiefer & Pulvermüller, 2012). Accessing these concepts as, for example, when retrieving word meanings involves a partial reactivation of the same brain processes that are active when experiencing the objects, situations or actions to which these concepts refer. This is also referred to as experiential simulation (Barsalou, 1999; Pecher &

Zwaan, 2005). These semantic effects occur within 100 - 200 ms after presentation of verbal stimuli, near-simultaneously to a range of psycholinguistic processes during comprehension (Pulvermüller, Shtyrov, & Hauk, 2009), and can therefore not be reduced to post-comprehension processes (Hoenig, Sim, Bochev, Herrnberger, & Kiefer, 2008). Furthermore, sensorimotor activations are modulated by context, allowing for a high degree of flexibility and fluency in the language comprehension system (Aravena et al., 2014; Hoenig et al., 2008).

There is ample evidence that experientially grounded meaning plays an essential role in conceptual knowledge (e.g., Binder & Desai, 2011) and language comprehension (for overviews see Bergen, 2015; Kaup et al., 2016; Meteyard et al., 2012; Pulvermüller, 2018).

Language-space associations

A particularly well-investigated domain of experiential grounding in language comprehension are language-space associations in the vertical dimension. Spatial locations do not by themselves form a natural category and there is no a priori thematic or associative link between objects sharing the same space within the upper or lower sphere (e.g., between 'kite', 'bird's nest' and 'crown' as objects typically found in the upper sphere of our world). Therefore, experiential traces of space seem particularly well suited to investigate the role of situational and experientially grounded meaning during language processing, as spatial locations can easily be inferred but are an implicit aspect of meaning. Due to reactivations of actual experiences during concept acquisition, processing nouns referring to objects with a typical location in space leads to an orientation of attention towards this location (e.g., Dudschig et al., 2012; Estes et al., 2008; Öttl et al., 2017). These reactivations of experiential traces of space are tied to simulations of contexts or events in which an object typically appears and cannot be deduced to abstract meaning features like UP or DOWN (Ostarek & Vigliocco, 2017). Furthermore, spatial cues linked to situations can facilitate the accessibility of words, as has been shown in an anagram solving task (Berndt et al., 2018). Most studies on language-space associations have focused on spatial compatibility effects where the dependent measure bears spatial characteristics, such as an up- or downward movement (Lachmair et al., 2011), thus investigating an effect of language on non-linguistic tasks. Further, some studies used non-linguistic cues and investigated whether this influenced concurrent language processing. For example, Lachmair and colleagues (2016) changed the body position of their participants between an upright or a head-down position. They found that participants remembered more up-words in the upright position and more down-words in the head-down position. In another study, vertical visual motion of dots on a screen had an impact on a lexical decision task when participants were presented with verbs denoting upward or downward movement like *rise* or *fall* (Dudschig et al., 2013; Meteyard et al., 2008). Thus, perception of motion can influence language comprehension (see also Kaschak et al., 2005), hinting at a link between visual and semantic processes.

Experiential grounding and language production

While experiential grounding in comprehension is well-investigated (see above), comparatively little is known about the potential role of experientially grounded meaning in language production, and it is unclear whether experiential traces are among the meaning factors that determine which lexical candidates are selected for articulation.

Two of the few studies suggesting that experientially grounded motor information may influence subsequent language production used a cyclic naming paradigm. In this paradigm, visually depicted actions were blocked according to their effector (hands/arms vs. feet) and an interference effect for naming action verbs was found (de Zubicaray et al., 2017; Hirschfeld & Zwitserlood, 2012). In a second experiment by Hirschfeld and Zwitserlood (2012) participants were asked to produce action verbs for depicted actions while executing a concurrent motor task. When the effector of a depicted action (e.g., foot for the activity of jumping) matched the effector which had to be used for the concurrent motor task, interference in naming was observed, too. However, according to Hirschfeld and Zwitserlood, the results are also compatible with the view, that abstract foot- or hand-related semantic features were co-activated by the movements, spreading to abstract effector-related concepts like e.g., PART OF THE LOWER EXTREMITIES, HAS TOES/FINGERS, USED FOR WALKING/MANIPULATING OBJECTS which then lead to competition between activated lexical nodes (see also Vigliocco et al., 2002). Therefore, they argue that their findings might not be interpreted as clear-cut evidence for a direct functional role of experientially grounded conceptual representations in language production.

Similar results have been obtained in other picture naming tasks. Investigating the motor domain, Witt and colleagues (2010) asked their participants to squeeze a ball in one hand, slowing down the naming of tools whose handles faced the squeezing hand compared to naming animals (but see Matheson et al., 2014). In an object naming task which was combined with a concurrent manual task, an increase in object naming errors was found which was related to the degree of experience subjects had in touching the depicted objects: for frequently manipulated objects naming was more difficult when the concurrent motion task engaged the hands in a way which would make interaction with the real object impossible (Yee et al., 2013; for similar results using rTMS see Pobric, Jefferies, & Lambon Ralph, 2010b). Furthermore, patients with motion-related neurological diseases, like Parkinson's, show increased difficulties in verb-naming tasks as the degree of motor content of the depicted actions increases (Herrera et al., 2012).

Asking participants to provide a verbal label for a given definition, Fargier, Montant, and Strijkers (2019) found that words which are strongly grounded in sensorimotor and/or emotional experiences are retrieved faster than words which are lesser-grounded, irrespective of their concreteness. These results seem to support the importance of experientially grounded meaning aspects for lexical retrieval.

Taken together, few studies have investigated experientially grounded meaning in language production. Among those, some have investigated the role of experiential meaning in conceptual representations in general, employing mainly naming tasks, but without directly focusing on language production (Matheson et al., 2014; Mulatti et al., 2014; Sixtus et al., 2018; Witt et al., 2010; Yee et al., 2013). Furthermore, other studies provide little or inconsistent evidence concerning the role of experiential traces in lexical selection. Firstly, it is still unclear whether the involvement of sensorimotor simulations during picture naming is necessary (de Zubicaray et al., 2017; Hirschfeld & Zwitserlood, 2012). Secondly, the activation of sensorimotor traces seems to be highly context specific (Ben-Haim et al., 2015; Matheson et al., 2014). Moreover, there is evidence for both facilitation and interference of lexical access when providing information boosting experiential simulations (de Zubicaray et al., 2017; Hirschfeld & Zwitserlood, 2012; Mulatti et al., 2014; Sixtus et al., 2018; Witt et al., 2010). This pattern mirrors the findings in language comprehension research, where both interference and facilitation effects have been reported. However, to date, a clear and encompassing theory for these patterns still seems to be missing (Ostarek & Huettig, 2019). Therefore, the role of experientially grounded meaning aspects during lexical selection remains unclear.

Combining experientially grounded meaning aspects with distributional semantics

So-called hybrid models are theories of semantic memory which integrate accounts of meaning based on experiential grounding with accounts based on distributional semantics. Theories of distributional semantics assume that the statistical regularities in natural languages are taken up by the cognitive system and are transferred into semantic representations which reflect the use of language (see below for more detail). According to the distributional hypothesis “you shall know a word by the company it keeps” (Firth, 1957), the meaning of a word can be deduced by the linguistic context in which it occurs. This idea has been implemented in different kinds of computational models quantifying meaning similarity between words by computing co-occurrence vectors (for an overview see e.g., Günther et al., 2019; Sahlgren, 2008; Wingfield & Connell, 2022). While implementations of distributional semantic models approximate human performance in many different tasks, they lack psychological plausibility as they can’t explain how concepts acquire meaning, which has also come to be known as the symbol grounding problem (Harnad, 1990; Searle, 1980). On the other hand, experiential accounts of meaning tend to disregard the importance of non-sensory and non-motoric sources of semantic knowledge.

Theories of distributional semantics and theories of embodiment or experiential grounding of semantics have often been treated as separate while a combination of these accounts in fact helps our understanding of semantic memory (Davis & Yee, 2021). Given that we learn concepts not only from direct sensory experience but also merely by being immersed in language given a sufficiently large directly-grounded vocabulary (Louwerse, 2018) it becomes evident that the often-conceived gap between language-based distributional models of semantics and experientially grounded accounts of meaning is more dichotomous than necessary. Language use as reflected in large text corpora captures many aspects of our bodily and sensory experiences as we use language to communicate about them (Durda, Buchanan, & Caron, 2009) and therefore, sensorimotor contingencies are not only part of our direct experience but are also mirrored in distributional language use (Zwaan & Madden, 2005). Furthermore, we are able to learn about bodily and sensory experiences merely by being exposed to linguistic descriptions of these without first-hand experience but still yielding typical effects of experiential re-activation (Günther, Dudschig, & Kaup, 2018; Günther et al., 2020) pointing to the fact that oral and written language can in fact serve as just another source of experience. These observations led to several calls for reconciling grounded and distributional accounts of meaning (Andrews, Frank, & Vigliocco, 2014; Davis & Yee, 2021).

In summary, language is used to communicate about the world and our experiences in the world and therefore it is not independent from it. Distributional semantics, which rely on the statistical regularities in language use, therefore often also contain information about sensorimotor experiences (Louwerse, 2011). However, the correspondence between our direct sensorimotor experiences of the physical world and the experiential information extracted from language use alone is not 1:1. There are meaning aspects which can only be inferred from one of these sources (for a detailed discussion of the relation between sensorimotor grounded meaning and distributional semantics see Günther et al., 2019) and at least part of our mental lexicon needs to be directly grounded (Vincent-Lamarre et al., 2015). This claim is backed up by increasing evidence that sensorimotor and distributional-linguistic meaning aspects are interacting but distinct types of knowledge.

For example, Carota and colleagues (2021) found a widely distributed network of active brain regions during silent reading. Importantly, activity in brain regions relevant for semantic selection and combinatorial semantic processes correlated with a distributional model of the stimulus set while cortical regions associated with sensorimotor processing responded more strongly to the experience-based characteristics of the stimulus set.

While it is acknowledged that insights into the nature of semantic representations – which have mostly been gained by investigating language comprehension – should be incorporated into language production research (Vinson et al., 2014), neither theories based on distributional language usage nor experientially grounded theories – or a combination of both – played an important role in the investigation of lexical selection processes. Only recently, Banks and colleagues (2021) asked participants to produce category members for given semantic categories. They found that both the order and the frequency of produced words can be predicted by measures of linguistic and sensorimotor similarity. These findings were also integrated into a computational model which performed most accurately with indirect spread of activation between categories and when sensorimotor and linguistic distributional aspects of meaning were accounted for. This is one of the first pieces of evidence suggesting that speakers make use of the experiential and linguistic contexts in which words occur and that they contribute separately when it comes to lexical selection.

However, an explicit integration of various aspects of meaning in language production models is still lacking (Abdel Rahman & Melinger, 2019; Vinson et al., 2014) and we know little about the role of distinct types of information from which word meaning can be learned (Louwerse, 2018; Vigliocco, Meteyard, Andrews, & Kousta, 2009).

The present study

This study was designed to test whether experientially grounded meaning aspects have an influence on which words we select when we prepare to speak and in how far they are influenced by distributional aspects of meaning. We combined the existing paradigms from the comprehension literature which show that physical visual stimulation has an influence on the processing of spatially connotated words (Berndt et al., 2018; Dudschig et al., 2013; Kaschak et al., 2005; Meteyard et al., 2008; Ostarek & Vigliocco, 2017) with the evidence for automatic re-activation of spatial meaning when processing up- and down-related words and sentences (Bergen et al., 2007; Dudschig et al., 2012; Estes et al., 2008; Lachmair et al., 2011; Lachmair, Dudschig, de la Vega, & Kaup, 2016; Ostarek, Ishag, Joosen, & Huettig, 2018; Öttl et al., 2017; Thornton, Loetscher, Yates, & Nicholls, 2013; Vogt et al., 2019). We developed a paradigm which enables us to investigate whether activations of language-space associations – for which there is ample evidence in language comprehension – can be found in language production, too. To this end, we employed a free production task and manipulated both the visual presentation mode (up- vs. downward movement of sentences) and the spatial content of the stimulus sentences (describing different locations in space). In contrast to previous studies investigating the duration of lexical selection processes (e.g., Hustá et al., 2021), we asked WHICH words are selected based on contexts that pose little or no semantic constraints. Participants were asked to complete written sentence fragments (e.g., <*You are strolling across the field and you see ...*>) by orally producing a noun of their choice. The fragments extended upwards or downwards from the center of the screen, with each word being presented above or below the previous word.

As visual input has been shown to influence the processing of words with spatial connotations, we assumed that visual stimulation also influences lexical-semantic processing during language production. We expected lexical choices for completing the sentence fragments to be influenced by visuo-spatial manipulation; that is, the location of the produced nouns should be predicted by the upward or downward movement of the sentence fragments. In other words, participants should complete a sentence like <*You are hiking through the forest and you see a…*> with a noun like *bird*, referring to an entity that is typically found in the sky, after having read an ascending sentence, and with a noun like *fox* after having read a descending sentence.

Additionally, we examined influences of the spatial location of the situation described by the sentences, investigating whether the typical location of the produced nouns can be predicted by the spatial connotations of the sentences. After sentences denoting situations which are perceived as occupying a higher physical space like <*You are in the mountains …*>, we expected nouns to refer to entities in the upper sphere and vice versa.

Moreover, we estimated the degree of semantic similarity between the produced noun and the noun in the sentence fragment using cosine values as a distributional measure of similarity (Günther et al., 2015). Semantic similarities are computed based on text corpora, and meaning relations of words that tend to occur in similar texts may capture different semantic relations as categorical, associative or thematic links (Durda et al., 2009). Therefore, we used the distributional measure of similarity to obtain an estimate of semantic relatedness that captures the traditionally investigated semantic relations known to influence lexical selection during language production in semantic context paradigms. By relating our experiential spatial manipulations to a measure of semantic relatedness, we addressed the question of how experientially grounded and linguistic distributional semantic meaning aspects relate to each other in a production task with given sentence contexts.

Experiment 1

Methods

Participants. We recruited 35 native German speakers using the institutes' participant pool Psychologischer Experimental Server Adlershof (PESA). The data of two participants was removed prior to analysis due to a high number of missing or invalid answers (less than 60 % of remaining trials). The final sample consisted of 33 participants (24 females, 18 - 33 years, $M_{\text{age}} = 25.82$, $SD_{\text{age}} = 4.56$) who provided written informed consent prior to participation. The study was conducted according to the principles expressed in the Declaration of Helsinki and was approved by the local Ethics Committee. Participants received either course credit or monetary compensation.

Stimuli. 90 German sentences like *<Du spazierst über das Feld und siehst eine . . . >* (English: *<You are strolling across the field and you see a . . . >*) or *<Du gehst zu der Haltestelle und siehst einen . . . >* (English: *<You are walking towards the bus stop and you see a . . . >*) were used as stimuli. All sentences had a similar structure and were incomplete. The 1st position of each sentence consisted of the personal pronoun *you.* At the 2nd position, 30 verbs of motion (of which nine were stative verbs, e.g., *walk, stroll, run, sit, stand*) were used; thus, each verb appeared in three different sentences. The 3rd position consisted of a local preposition, followed by a definite article at the 4th position. The 5th position constituted a noun containing the relevant information regarding the scene of the described event. Nouns were only used once and referred to an individual's destination or places where a person can move around (e.g., *street, field, bus stop, forest, lake, train station*). The sentences continued with the conjunction *and* at the 6th position and a verb of perception at the 7th position (*see, spot, discover*), each repeated 30 times across all sentences. At the 8th position, there was an indefinite article. As accusative articles in German signify gender, we counterbalanced the distribution of neutral, female and male articles over six experimental lists, assuring that each sentence was paired with each article equally often across participants and experimental conditions. After the indefinite article, the sentences ended with an ellipsis to prompt participants to complete the sentence. We ensured that a wide range of endings was possible for each sentence, i. e. sentences were not constraining as for example in cloze paradigms (Block & Baldwin, 2010). We used 40 filler sentences with a similar structure as our experimental sentences. The ending for some fillers was intended to be more easily predictable in order to make the task easier for participants. Six additional sentences were used in practice trials.

Sentence spatial location. Before starting the main experiment, we conducted an online rating of the spatial locations of our sentences using the platform https://www.soscisurvey.de. Nine voluntary participants who did not take part in the main experiment (6 females, 22 - 67 years, $M_{\text{age}} = 31.78$, $SD_{\text{age}} = 13.63$) indicated on a 7-point Likert scale where the places denoted by the noun at the 5th sentence position are in space (see below for more information on spatial ratings). These values served as a measure of the spatial location of the scenes denoted by the sentences. These values were later added into our analysis in order to analyze the impact of the sentence location on the choice of a suitable sentence ending. A list of all sentences with their respective spatial location values are presented in Appendix A.

Procedure. Before starting the experiment, we told our participants that we were investigating language processing of speakers with different native languages (Arabic, Chinese and German) as a cover story. We deemed it common knowledge that Arabic and Chinese differ from German regarding reading direction and wanted to keep participants from wondering why stimuli were presented in an unusual reading direction in order to minimize the risk of participant's guessing the aim of the task.

Participants were seated in a dimly lit room approximately 70 cm in front of a computer screen with a resolution of 1280 x 1024 pixels. Sentences were displayed consecutively in Rapid Serial Visual Presentation mode using Presentation® software (Version 17, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com). The experiment started with a practice block consisting of six sentences. During the experiment, participants were able to take small breaks after blocks of fifteen sentences.

Each sentence was presented once during the experiment and participants saw each sentence either in ascending or descending presentation direction. Filler sentences were presented in the same way. Within each list, test sentences and fillers were presented in a random order.

We presented each word for 300 ms in black on a grey background in Arial 24pt font. Each trial started with a fixation cross appearing in the center of the screen for 500 ms. Then the first word appeared in the center of the screen. The following word replaced the previous word and appeared either 35 pixels higher or lower than its predecessor. Thus, the position of the three dots was 315 pixels above or below the screen center and 197 pixels apart from the edge of the screen. The dots remained on screen for 4000 ms; afterwards there was a blank screen for 2000 ms before the next trial started. Participants were instructed to read the sentences silently and to complete the sentences with a suitable noun as spontaneously and quickly as possible as soon as the ellipsis appeared. Participants were asked to orally produce only one word in each trial and to avoid repeating the same noun several times throughout the experiment. We recorded answers given in the time frame of 6000 ms after the ellipsis appeared (see Figure 1 for illustration of a trial sequence). The experimenter monitored the experiment from another room and immediately noted the answers.

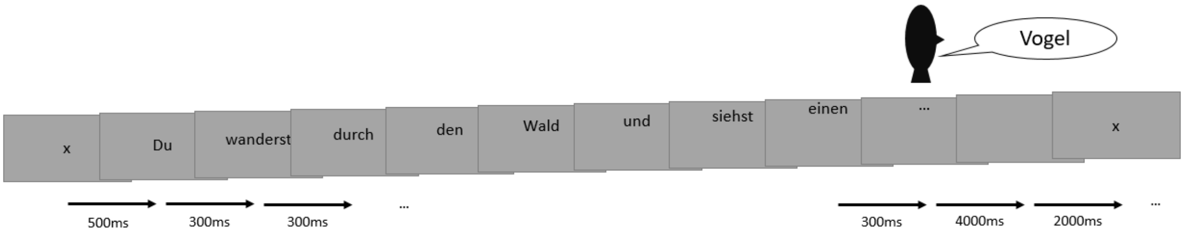

Figure 1. Trial sequence with an example of an upward moving sentence (English: <*You are walking through the forest and you see . . .*>) and a participant producing the noun *bird.* Note that screen position was fixed in the experiment and only moves upwards in the figure for illustrative purposes.

Rating of spatial attributes of produced nouns. In a second step, after running the sentence completion study, spatial attributes for the produced nouns were obtained to assess whether the choice of the produced nouns had been influenced by the experimental manipulation. To this end, the produced nouns entered a rating study. Nouns uttered by several participants were included only once (e.g., several people used the word *bird*, albeit some used it in different contexts throughout the experiment). Nouns which presumably have the same referent –

but where participants used different lexemes to convey a comparable meaning – entered the rating in all the forms which had been produced during the experiment (e.g., *Schiffsanleger* vs. *Bootsanleger*, English: roughly *jetty* vs. *pier*). In case of ambiguous nouns, a short description of their lexical meaning was added. For example, as could be inferred from the context, the word *Sirene* had not been intended to refer to the English *siren*, but to the mythological figure of a mermaid. Therefore, raters saw this noun as *Sirene (Fabelwesen)*, English: *siren (mythological figure)*. The complete set of produced words was reduced to a set of 1056 rated words, implemented similarly to Díez-Álamo, Diéz, Wojcik, Alonso, and Fernandez (2018) and Scott, Keitel, Becirspahic, Yao, and Sereno (2019). In order to reduce the time for the rating for each rater, each rater only saw a subset of the total word set. To this end, the produced nouns were randomized and then distributed on 21 questionnaires, with the first questionnaire containing words 1-352, the second questionnaire containing words 51-402, the third questionnaire containing words 101-452, etc., thereby ensuring that the questionnaires were representative of the whole set of produced nouns. Additionally, 20 control words were added to each questionnaire. These control words had not been produced in the sentence completion study but were taken from an unpublished set of spatial ratings for German nouns, spanning the entire range of locations where entities can be encountered, from very high up (*Sternschnuppe*, English: *falling start*) to very low (*U-Boot*, English: *submarine*). The questionnaires were administered using the online platform https://www.soscisurvey.de. In each trial, a target word was selected randomly and displayed with a vertical 7-point rating scale ranging from 'up' to 'down' (with 'centrally' at the midpoint) below it. Additionally, participants could skip the rating of a word in case the spatial property could not be judged. By clicking one of the points on the scale, participants had to judge where the object referred to by the noun can typically be found. The approximate time to complete the rating was 20 minutes. In total, 37 voluntary participants who did not take part in the production experiment (22 females, 23 - 59 years, $M_{\text{age}} = 35.76$, $SD_{\text{age}} = 9.61$) were randomly assigned to one of the questionnaires. The procedure of assigning different questionnaires to participants ensured that each target word received ratings from at least 9 subjects. Assuming that ratings for control words whose spatial location was based on previous rating data are an indicator of subject's compliance, intra-class correlation between the previous rating data and each rater was computed using the function ICC from the R-package psych (Revelle, 2018), as suggested by Hallgren (2012) and Trevethan (2017). The agreement with the existing mean rating values for the control words was $ICC(3{,}1) \geq .72$ for all raters and the intra-class correlation between all subjects was excellent following the criteria of (Fleiss, 1986), $ICC(3{,}1) = .83$. Therefore, none of the ratings were excluded, and ratings for all but the control nouns were averaged across raters, yielding one rating value for each distinct noun. This served as an indicator of the spatial location of the entity denoted by that noun. Ratings were merged with the data from the sentence completion study so that for each trial a mean spatial rating serving as an indicator of the spatial location of the produced noun was obtained.

Data Analysis

Data was analyzed using the free statistics software R Version 3.6.1 (R Core Team, 2017). The data set consisted of 2970 data points (33 participants completing 90 sentences). Missing trials in which participants did not produce a noun were excluded (12.2 % of all trials). Afterwards, erroneous trials (incomplete, unintelligible, and nonsensical answers as well as utterances which consisted of more than one word or in which participants simply repeated the noun of the sentence which they had read) were excluded from further analysis (2.9 % of all trials). Additionally, eight trials had to be excluded for missing spatial rating values due to experimenter error. Trials in which participants produced nouns whose gender did not match the gender required by the article were not excluded from analysis as some participants seemed to have ignored the gender of the article. There were many instances of masculine nouns being produced after neutral articles, which is incorrect from a grammatical point of view. However, the masculine accusative article *einen* is typically shortened to *ein* in colloquial speech, equaling the neutral article. Thus, it cannot be safely concluded that participants ignored the gender of the article in those cases, as they might have silently pronounced the written sentences before giving an answer aloud. The phonological similarity of *einen* and *ein* in spoken German might have led them to produce nouns of both neutral and male gender, respectively. In total, a set of 2515 utterances remained for statistical analysis.

In order to assess the influence of the experimental manipulation on the spatial properties of the produced nouns, a linear mixed model was computed with the packages lme4 (Bates, Mächler, Bolker, & Walker, 2015) and lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017). We started with a maximal model containing interactions between the fixed predictors presentation direction and centered spatial location values for the nouns of the sentence fragments, as well as by-subject and by-item random intercepts and slopes. Sliding difference contrasts were applied for the predictor presentation direction. Random effects were simplified in case of singular fit or convergence problems, resulting in the final model containing by-subject and by-item random intercepts only. Using model comparison, this model was compared to one containing additive fixed effects for presentation direction and centered spatial location values. We report *beta*-estimates together with a 95 % confidence interval estimated with the Wald method, as well as *t*- and *p*-values.

Results

Numerically, there was almost no difference in mean spatial ratings between nouns produced after ascending vs. descending sentences ($M_{\text{up}} = 3.577$, $M_{\text{down}} = 3.582$). This finding was corroborated using a linear mixed model containing additive effects for presentation direction and spatial location values for stimuli, which explained the data better than a model containing interactions ($Chi^2(1) = .971$). There was no main effect for presentation direction ($\beta = -0.01[-0.10; 0.07]$, $t =$ -0.26, $p = .80$), but there was a significant main effect of sentence spatial location ($\beta = 0.20[0.11; 0.28]$, $t = 4.69$, $p < .001$), indicating that the spatial locations of the situations presented by the sentence fragments influenced the spatial attributes of the produced nouns, see Figure 2.[1]

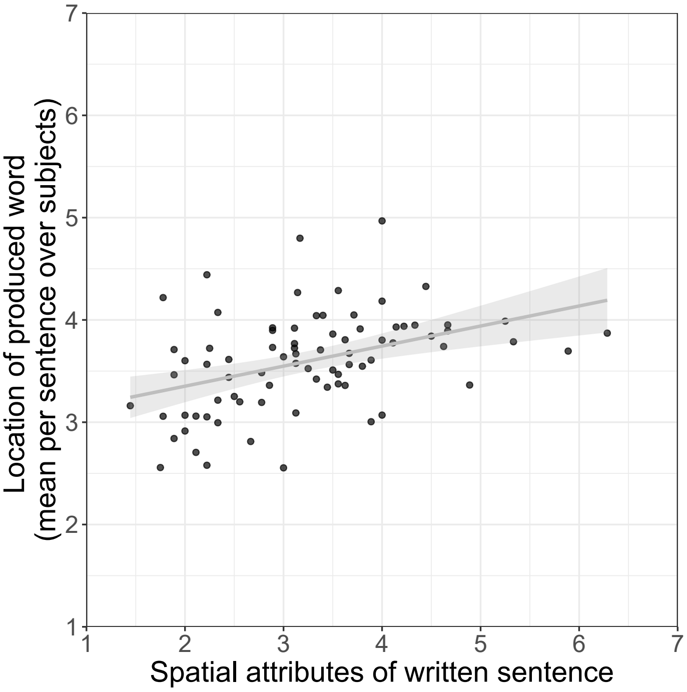

Figure 2. The spatial location of nouns in the sentence fragments predicts the location of the noun referents chosen as suitable sentence endings. The line depicts the effect as estimated in the linear models, the dots represent mean spatial values of the produced words for each sentence fragment respectively. Spatial locations of the entities referred to with the produced nouns were rated on a scale ranging from 1 (down) to 7 (up) after the experiment. Spatial locations of nouns in the sentence were rated on a scale ranging from 1 (down) to 7 (up) before the experiment. For illustrative purposes, sentence noun spatial locations are not centered.

To gain further insights into the relation between this effect and traditionally investigated semantic measures known to affect conceptual-semantic processing during language production, we included a distributional measure of semantic similarity between the nouns in the presented sentences and the produced nouns as a covariate in the analysis. We used the semantic space dewak100k_cbow (Günther et al., 2015) built from the deWaC-corpus by using the cbow algorithm as implemented in the word2vec model (Mikolov, Chen, Corrado, & Dean, 2013). The deWaC corpus is a 1.7 billion word corpus constructed from the Web limiting the crawl to the .de domain and using medium-frequency words from the Süddeutsche Zeitung corpus and

[1] We also collected ratings for the spatial locations of the whole sentence fragments. Participants rated the spatial position of these sentences independently of how they might be completed. Entering these sentence spatial locations as predictors yielded the same effect, with no effect for presentation direction and even stronger increases in slope for each level of spatial location from 1 (down) to 7 (up), (β = 0.38[0.27; 0.49], t = 6.78, p < .001).

basic German vocabulary lists as seeds (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009). Cosine values were computed for each pair of 5th sentence position and produced nouns using the package LSAfun (Günther et al., 2015). These cosine values serve as an indicator of semantic similarity, with higher values indicating that the two respective words more often occur together in similar contexts than others and have a highly similar meaning.

Because not all words were included in the used corpus, cosine values could not be computed for all cases. Furthermore, trials with cosine values less than zero were not used for subsequent analyses as these cosine values cannot be interpreted in a meaningful way (Günther et al., 2015). Thus, the reduced data set with similarity measures consisted of 1904 out of 2515 total nouns which had been used for the first linear mixed model analysis.

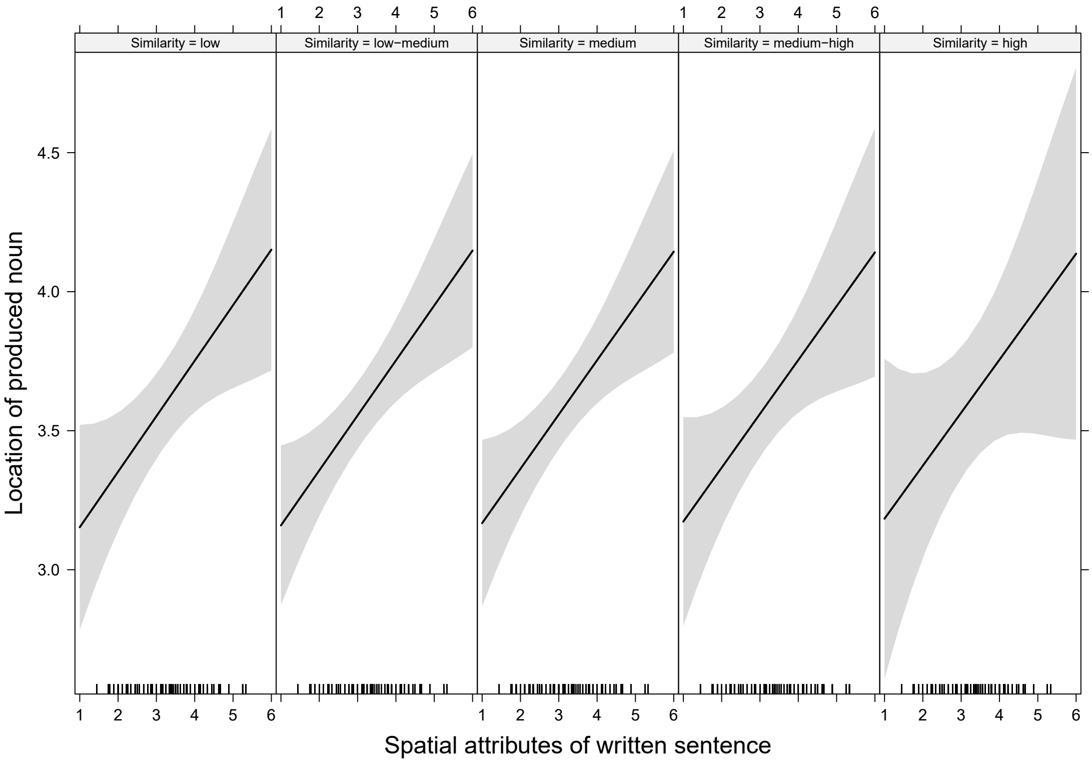

Figure 3. Effect plot showing that increasing degrees of semantic similarity between the noun in the presented sentence and the produced noun did not influence the spatial attributes of the presented sentences. Higher values for location of the produced noun as well as for spatial attributes of the written sentences indicate a higher localization in space. Small ticks above the x-axis mark the distribution of the set of sentences regarding their spatial properties. For illustrative purposes, sentence spatial locations are not centered. The continuous predictor of similarity was split into five points of equal distance. Low and high similarity refer to the lowest vs. highest cosine values obtained in this study; they are used as descriptive labels while no pre-defined level of degrees of semantic similarity regarding cosine values exists.

Centered cosine values were entered into the linear model as an additional predictor with main effects for direction and an interaction between cosine values and centered sentence spatial location values, as well as random intercepts for items and subjects. There was no interaction between semantic similarity as indexed by cosine values and sentence spatial location values ($\beta =$ 0.08[-0.35; 0.50], $t = 0.36$, $p = .72$). Thus, the effect of sentence spatial location on spatial properties of the produced nouns cannot be explained by similarity, see Figure 3.

Again, there was no effect of direction but a significant main effect of sentence spatial location (β = 0.20[0.10; 0.29], t = 4.07, $p < .001$). For similar results obtained with the semantic space de_wiki, see supplement S1.

Discussion

There was no influence of the manipulation of presentation direction on the spatial properties of the produced nouns. Therefore, our hypothesis that visual spatial manipulation in the form of a physical spatial cue affects lexical selection was not confirmed. However, there was an influence of the experientially driven meaning dimension LOCATION IN SPACE on the choice of nouns. When considering the typical spatial location of the situations described by the sentence fragments, the spatial properties of the produced nouns could be predicted. Thus, the more a sentence referred to a situation in the upper or lower domain of the world, the higher up or lower down the referents of the produced nouns were located. For example, after the sentence <*You lean at the window and you see a...*> which had been rated as being found in the upper sphere, the nouns people produced tended to be more upward related like *bird* or *rainbow* as when people completed sentences like <*You jump over the tree trunk and you see a...*> which had been rated as being in the lower sphere and where people were more likely to produce words as *rainworm* or *hole* which are also more downward related in comparison to upward related words like *bird's nest.* This might demonstrate that experiential traces of space are reactivated during language processing and influence subsequent lexical selection. We will discuss this interpretation in the General Discussion section.

However, many participants reported that the task was difficult for them, reflecting the high number of lost and invalid trials (15.1 %, see Methods). We had aimed to prevent participants from preparing a possible answer prior to reaching the end of the sentence by also presenting an indefinite article. As German articles determine gender in the accusative case, participants had to wait until they read the article before a lexical choice could be made. Thereby, we wanted to maximize the impact of the visual manipulation and prevent participants from preparing their answer in advance. However, this manipulation made it difficult to come up with suitable nouns, as time for completing a sentence was limited and led to omissions, neglecting the case of the article, or – as in about 20 % of all trials – naming a person. This was a wide-spread strategy in order to fulfil the gender requirements of the article. For instance, participants could say *Polizist* in case of the male article *einen* (English: *policeman*) and *Polizistin* in case of the female article *eine* (English: *policewoman*). However, naming a person is not informative about the spatial attributes of a noun, as persons are usually found in the central plane and occupy the same space as a person experiencing the situation described by the sentence. The large number of trials in which a noun referring to a person was produced may have reduced the impact of the movement manipulation. In Experiment 2, we therefore presented sentence fragments with no articles.

Experiment 2

Experiment 2 was a preregistered study using the Open Science Framework (https://osf.io/se6a3/?view_only=f666716d3b8f47228017b9dadc6e2950). It was designed to replicate the findings of Experiment 1. To reduce task complexity and trial loss, words were presented for slightly longer and sentences didn't end with an article. Thus, participants were asked to produce a determiner noun phrase and were not restricted in their selection of suitable nouns regarding gender. Furthermore, we introduced a baseline condition in which sentences were presented in the center of the screen. We also improved the stimulus set by balancing the sentence spatial location values of the sentence fragments across the different presentation directions. Additionally, the number of participants was increased to enhance the chances of detecting even small effects of the movement manipulation on lexical selection.

Methods

Only those aspects differing from the first experiment will be described below.

Participants. We recruited 78 native German speakers.[2] Data of two participants was removed prior to analysis as their German language proficiency was limited despite reporting being native speakers. Furthermore, the data of four other participants was excluded due to not following the instructions (n = 2) or a high number of missing or invalid answers (n = 2). The final sample consisted of 72 participants (49 females, 18 - 35 years, $M_{\text{age}} = 25.60$, $SD_{\text{age}} = 4.93$). Participants provided written informed consent prior to participation. The study was conducted according to the principles expressed in the Declaration of Helsinki and was approved by the local Ethics Committee. Participants received either course credit or monetary compensation.

Stimuli. 60 German sentences with a similar structure as in the first experiment were used as stimuli, e.g., <*Du läufst zum Feld und siehst . . .*> (English: <*You are walking towards the field and you see . . .*>). Compared to the first experiment, only four verbs of motion (*stand, walk, go, enter*) and the verb *be* were used at the 2nd position, appearing equally often across the full set of sentences. The 3rd position consisted of a definitive article or a local preposition contracted with a definite determiner (e.g., *am* – English: *at the*, *zur* – English: *towards the*). At the 4th position, a noun conveying the relevant information about the location at which the scene happened was used. The sentences finished with the conjunction *and* at the 5th

[2] Sample size for Experiment 2 was based on preliminary results of Experiment 1 hinging on spatial rating data for the complete set of produced nouns stemming from three subjects only. This analysis yielded a significant effect of presentation direction for cases with high semantic similarity between sentence noun and produced noun. Based upon this finding, sample size for Experiment 2 was estimated (see preregistration). Even though the preliminary results turned out the be spurious after completing the spatial rating with 37 raters, we ran the second experiment with the originally planned sample size.

position and a verb of perception at the 6th position. The sentence display terminated with an ellipsis (<...>), serving as a prompt for the participants to complete the sentence with a suitable noun. Furthermore, we constructed 24 filler trials with a similar structure and the same number of words as the sentences. Six additional sentences were used in practice trials.

Sentence spatial location. Before starting the main experiment, we conducted an online rating of the spatial location of our sentences using the platform https://www.soscisurvey.de. Fifteen voluntary participants who did not take part in the main experiment (8 females, 27 − 71 years, $M_{\text{age}} = 34.00$, $SD_{\text{age}} = 10.76$) indicated on a 9-point Likert-scale where the places denoted by the noun at the 4th sentence position are in space.[3] Apart from adding them as predictors into our analysis, the sentence spatial location values were used to construct experimental lists. All sentences with their respective sentence spatial location values are presented in Appendix B.

Procedure. Mean rating values were computed for each of the sentence nouns ranging from 2.4 (*Kanal*, English: *canal*) to 7.7 (*Aussichtspunkt*, English: *vantage point*). Afterwards, three sentences with nouns of a similar mean rating value were combined into a triplet with the goal of minimizing the difference in mean rating values between nouns in the same triplet. The resulting difference was 0.31 or less, with a mean difference of 0.08 between the nouns of an adjacent sentence pair. Each participant read each sentence fragment from each triplet. Each participant saw each sentence of a triplet only once in one of the three presentation directions. Presentation direction and triplets were counterbalanced over nine lists so that every sentence was presented equally often in the same direction across participants. This resulted in every participant reading sentence fragments with similar spatial location values in each condition. Thereby, we controlled potential impacts of spatial locations on our stimuli with regard to the manipulation of presentation direction.

[3] A 9-point Likert scale for prerating of experimental items for Experiment 2 was chosen as we assumed that a wider range would better pick up on the big real word differences in spatial locations. However, as suggested in several methodological papers, 7-point Likert scales seem to be yielding the best reliability (Cicchetti, Shoinralter, & Tyrer, 1985; Finn, 1972; Oaster, 1989; Ramsay, 1973), we again chose a 7-point Likert scale for the rating of the produced nouns as this also enables direct comparison of results from Experiment 1 and 2.

Each trial started with a fixation cross presented for 500 ms, after which each word was presented for 400 ms. Each sentence ended with an ellipsis, serving as an indicator that participants should complete the sentence. After 4 sec, a circle was presented in the center of the screen. Then participants could start the next trial in a self-paced manner by pressing the space bar on the keyboard. Responses were recorded within a window of 6 sec after the appearance of the ellipsis. The first word of a sentence always appeared in the center of the screen. The following word appeared either at the same position, 47 pixels above or below that original position, or replaced the previous word at the center. The position of the ellipsis was 329 pixels above or below screen center and 183 pixels apart from the edge of the screen. Participants were instructed to read each sentence fragment and to spontaneously produce a noun to end the sentence as quickly as possible and, if necessary, with an appropriate determiner. We asked them to not complete the sentences by repeating parts of the sentence or by producing nouns describing a person (e.g., *a woman, a bus driver, a neighbor*). The experimenter monitored the experiment in the same room behind a folding screen, noting participants' answers on a sheet of paper.

Rating of spatial attributes of produced nouns. In order to minimize the total number of words entering the rating, nouns with the same – or very similar – referents entered the rating only in one form, assuming that the spatial properties of the referents of these almost synonymous words would be the same. Additionally, nouns produced several times by different participants entered the rating only once.[4] Thus, the total number of produced nouns was reduced to 915 words divided into 10 questionnaires, each containing 457 or 458 words as well as

[4] While 35 participants took part in Experiment 1 with 90 stimulus sentence, 72 participants took part in Experiment 2 with 60 stimulus sentences. Thereby, the total amount of potentially produced words which had to be rated already went up by ca. 25 %. Furthermore, several of the raters in Experiment 1 gave us the feedback that they found it annoying to rate spatial locations of seemingly equivalent objects. Our decision to only let people rate one version of near-synonyms thereby served both the purpose of making raters more willing to cooperate and of reducing the amount of words to be rated and thereby reducing the time needed for doing the rating and/or the amount of raters. For Experiment 2, two experimenters decided together whether two produced words could be handled as synonyms for the rating and in case of doubt both versions were kept in the rating. For this decision, we always kept in mind whether two words would refer to the same type of object and whether the use of one word version vs. the other could potentially have an impact on how far up or down in the world other people might perceive the word's referents.

20 additional control words. Data from one rater was excluded prior to computing mean spatial ratings, as the intra class correlation coefficient with the control words – which was only in the fair range (ICC(3,1) = .58) – indicated that this participant did not follow the instructions. Therefore, data from 30 participants (19 females, 18 - 76 years, $M_{\text{age}} = 33.57$, $SD_{\text{age}} = 13.82$) was used to compute mean spatial rating values, with each target word having been rated by at least 13 subjects. Mean spatial ratings were merged with the data from the sentence completion study in order to obtain a numeric indicator of the spatial location of each produced word.

Data Analysis

The data set consisted of 4320 data points (72 participants completing 60 sentences). Trials in which participants didn't produce a noun (2.4 % of all trials), as well as erroneous trials (1.4 % of all trials) and trials in which participants produced a noun describing a person, were excluded (4.2 % of all trials). Additionally, 13 trials for which no spatial ratings were obtained due to experimenter error were excluded from further analysis. Based on our preregistered criteria, all trials including the sentence <*You are at the harbor and you see...*> were excluded from analysis because more than 50 % of the participants chose the same noun to complete the sentence. After preprocessing, the data set consisted of 3893 nouns.

Like Experiment 1 and based on the preregistered analysis plan, we analyzed the data with a maximal model containing interactions between the fixed predictors presentation direction and centered spatial location values for the nouns of the sentence fragments as well as by-subject and by-item random intercepts and slopes. Sliding difference contrasts were applied for the predictor presentation direction (three levels: descending, central, ascending). Random effects were simplified in case of singular fit or convergence problems which resulted in the final model containing by-subject and by-item random intercepts only. Using model comparison, this model was compared to one containing additive fixed effects for presentation direction and centered spatial location values.

Results

As in Experiment 1, a model containing additive effects for presentation direction and spatial location values for stimuli explained the data best ($Chi^2(2) = .876$). Contrary to our hypothesis, nouns produced after ascending sentences were located lower in space ($M_{up} = 3.64$) than nouns produced after sentences with unchanging position ($M_{\text{central}} = 3.79$), resulting in a significant main effect for the contrast of ascending vs. central presentation direction in the linear mixed model (β = -0.15[-0.24; -0.07], t = -3.51, $p < .001$). There was no significant main effect for the contrast of descending vs. central presentation direction (β = 0.07[-0.02; 0.15], t = 1.50, p = .133), see Figure 4.

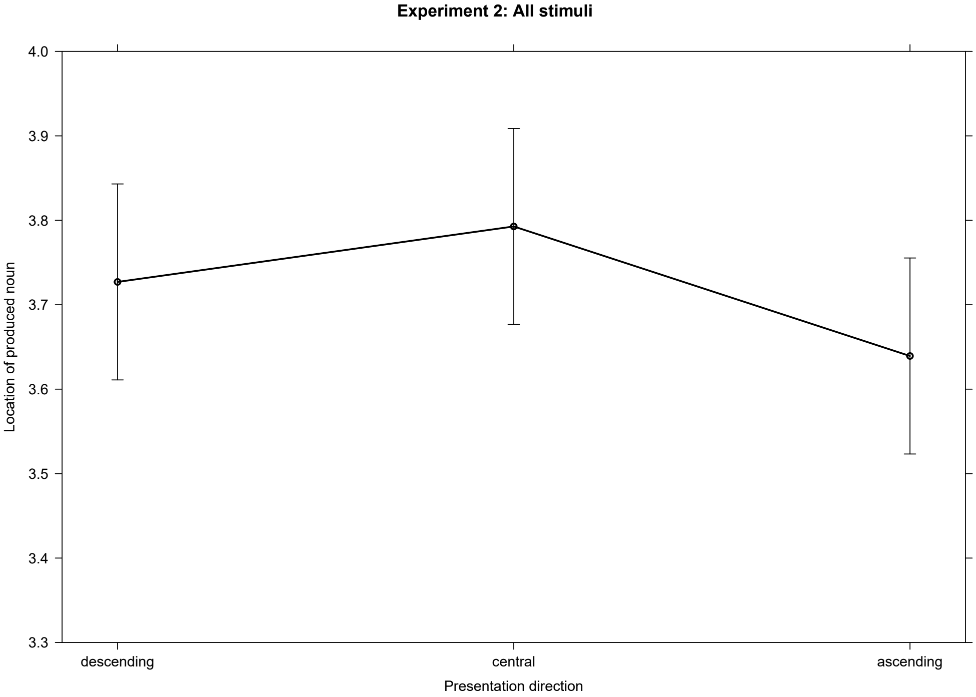

Figure 4. Estimated means and 95 % confidence band for spatial locations of the produced nouns depending on the presentation direction of sentence fragments in Experiment 2.

Furthermore, and converging with results from Experiment 1, there was a significant effect for sentence spatial location values (β = 0.28[0.19; 0.37], t = 6.32, $p < .001$), indicating that the locations of the sentence fragments influenced the spatial attributes of the produced nouns, see Figure 5.

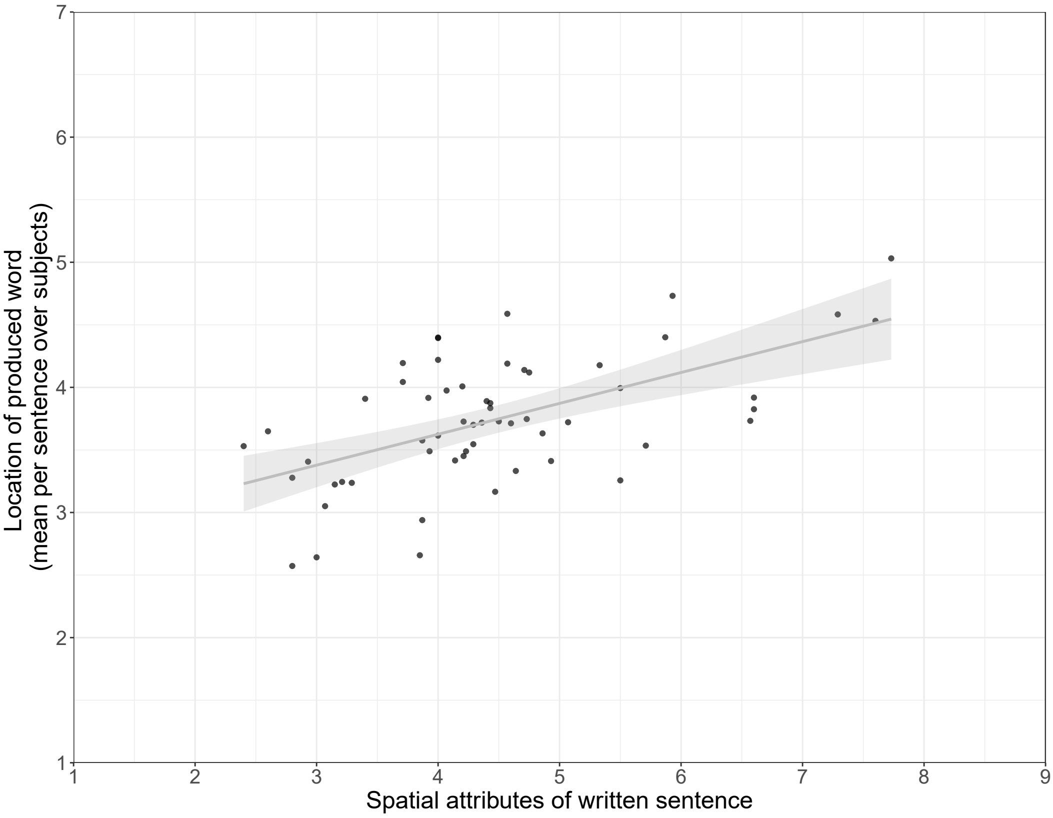

Figure 5. The spatial location of nouns in the sentence fragments predicts the location of the noun referents chosen as suitable sentence endings. The line depicts the effect as estimated in the linear models, the dots represent mean spatial values of the produced words for each sentence fragment respectively. Spatial locations of the entities referred to with the produced nouns were rated on a scale ranging from 1 (down) to 7 (up) after the experiment. Spatial locations of nouns in the sentence were rated on a scale ranging from 1 (down) to 9 (up) before the experiment. For illustrative purposes, sentence noun spatial locations are not centered.

For comparison with Experiment 1, an additional linear model was fitted post-hoc to allow for a direct comparison of ascending vs. descending presentation direction. Nouns produced after ascending sentences were located lower in space ($M_{up} = 3.64$) than nouns produced after descending sentences ($M_{\text{down}} = 3.73$), as demonstrated by a significant main effect for the contrast of descending vs. ascending presentation direction in the linear mixed model ($\beta = -0.09[-0.17; 0.00], t = -2.01, p = .045$).

The differential outcomes of Experiment 1 and 2 were further investigated by comparing results from the subset of sentences with overlapping noun use between experiments (27 out of 60 sentences). For the subset of sentences from Experiment 2, linear mixed models were specified as above without random intercepts for subjects due to singular fit. This again resulted in a significant difference between ascending and central presentation direction ($\beta = -0.18[-0.32; -0.04]$, $t = -2.47$, $p = .014$), as well as a marginally significant difference between central and descending presentation direction ($\beta = 0.14[0.00; 0.28]$, $t = 1.95$, $p = .051$) and a significant effect for sentence spatial location values ($\beta = 0.31[0.16; 0.46]$, $t = 4.16$, $p < .001$). In comparing ascending and descending presentation direction directly, no significant difference between ascending and descending presentation direction was obtained (β = -0.03[-0.18; 0.10], t = -0.52, p = .605), see Figure 6. Therefore, the difference between ascending and descending presentation direction when analyzing the complete data set of Experiment 2 seems to hinge on items exclusively used in Experiment 2.

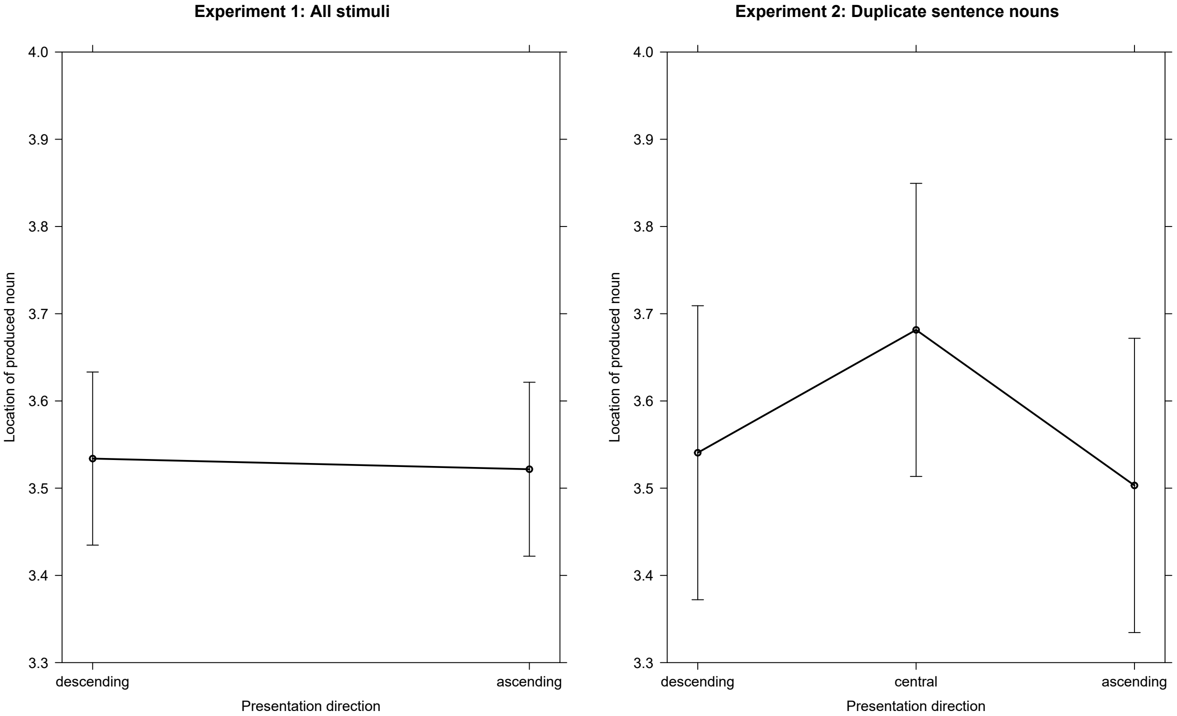

Figure 6. Estimated means of spatial properties of produced nouns depending on the presentation direction of the presented sentences and 95 % confidence intervals. There is no statistical difference between ascending and descending presentation direction in Experiment 1 (left panel) and when investigating the 27 overlapping stimuli from both experiments (right panel).

As in Experiment 1, semantic similarity measures were included in the model to test if the pattern in our data was influenced by the semantic similarity between the content in the displayed sentence and the produced noun. Therefore, cosine values were computed for each pair of nouns (one at the 4th sentence position and the one being produced) in the respective sentence using the semantic space dewak100k_cbow. Cosine values could be computed for 3570 trials out of 3893 which had entered statistical analysis. Centered cosine values were added to the linear model as an additional predictor with a main effect for direction and an interaction between cosine values and sentence spatial location values, as well as random intercepts for items and subjects. Again, there was a significant main effect of sentence spatial location ($\beta = 0.27[0.18; 0.36], t = 5.79, p < .001$) as well as a significant difference between central and ascending presentation direction (β = -0.10[-0.18; -0.01], t = -2.15, p = .032). Additionally, there was a significant interaction between semantic similarity – as indexed by cosine values – and sentence spatial location values (β = 0.54[0.26; 0.81], t = 3.86, p < .001), indicating that the effect of sentence spatial characteristics was influenced by the degree of semantic similarity between the sentence noun and the produced noun, see Figure 7. For similar results obtained by using the corpus de_wiki see Supplement S2.

We further investigated this interaction by splitting the range of obtained cosine similarities in between the highest and lowest similarity values in equally distant ranges. Then, we explored whether the effect of spatial characteristics of sentences on the produced nouns is contingent on a certain level of semantic similarity or if it exists across the entire range of semantic similarities.[5] Taking all trials from each level of similarity (low, low-medium, medium-high, high) into account, separate linear mixed models with the fixed predictors presentation direction, centered spatial location and random intercepts for items – as well as subjects – were computed. In case of singular fit, random effect structures were simplified. As shown in Table 1, the effect of spatial characteristics of sentence locations on spatial locations of produced nouns

[5] We deviated from the preregistered analysis plan of analyzing the impact of semantic similarity in bins of 10 %-percentiles for two reasons. Firstly, while cutting the whole set of similarities in percentile-bins would have permitted to run analyses with the same number of trials, it would have resulted in unequally spaced bins across the range of semantic similarity with the outer percentiles spanning a relatively large range of similarity values which is not informative with regard to the hypotheses. Secondly, splitting the whole set of similarities in ten equally spaced ranges would have resulted in outer bins not containing enough trials to run statistical analyses. Therefore, we resorted to splitting up the whole range of obtained similarity values using five levels.

is significant for each level of semantic similarity and gets more pronounced with higher degrees of semantic similarity between sentence nouns and produced nouns.

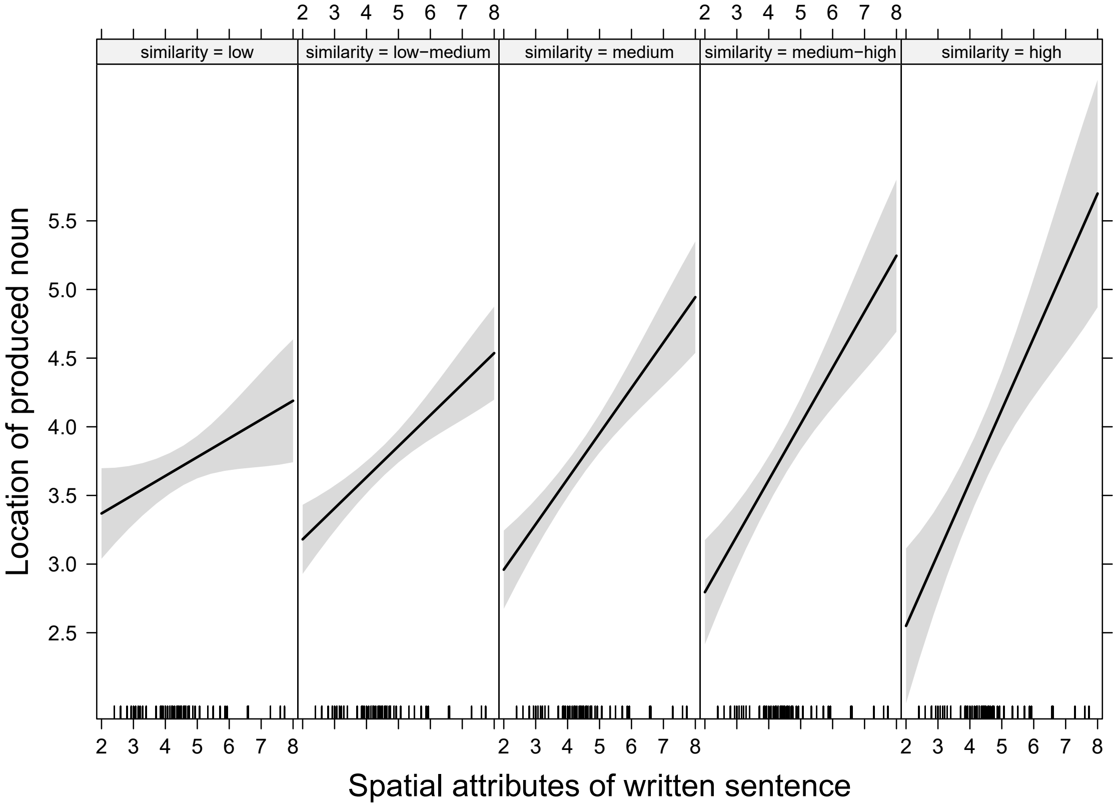

Figure 7. Effect plot showing more pronounced influences of the spatial attributes of the presented sentences on the spatial characteristics of the produced nouns for increasing degrees of semantic similarity between the noun in the presented sentence and the produced noun was more pronounced. Higher values for location of the produced noun – as well as for spatial attributes of the written sentences – indicate a higher localization in space. Small ticks above the x-axis mark the spatial property distribution of the set of sentences. For illustrative purposes, sentence spatial locations are not centered as they were in the analysis. The continuous predictor of similarity was split into five points of equal distance. Low and high similarity refer to the lowest vs. highest cosine values obtained in this study. They are used as descriptive labels, while no pre-defined level of degrees of semantic similarity with regard to cosine values exists.

Table 1

LMM statistics for the influence of presentation direction and sentence spatial location on the spatial properties of the produced nouns for different degrees of semantic similarity between the sentence noun and the produced noun based on centered cosine values. To illustrate the range of cosine values, the non-centered equivalents of the cosine values on which analyses were based are given together with example pairs from the data set consisting of a noun which had been part of the presented sentence and two exemplars of nouns produced after these sentences.

Similarity Range	low 0 < cos ≤ 0.216	low-medium 0.216 < cos ≤ 0.366	medium-high 0.366 < cos ≤ 0.516	high 0.516 < cos
Examples	tree – lady bug, picture house – stars, swing lake – animal, towel	tree – bicycle, swing house – fence, chair lake – inflatable mattress, fishing rod	tree – bird, squirrel house – garage, window lake – pier, water lily	tree – branch, leaves house – hut, garden lake – boat, shore
	n = 856	n = 1454 no random intercept for subjects	n = 965 no random intercepts for subjects	n = 295 no random intercepts for subjects

Variable	*b*	*CI*	*t*	*p*	*b*	*CI*	*t*	*p*	*b*	*CI*	*t*	*p*	*b*	*CI*	*t*	*p*
Intercept	3.66	[3.51;3.80]	49.27	<.001	3.72	[3.59;3.85]	55.48	<.001	3.69	[3.50;3.88]	38.04	<.001	3.70	[3.41;3.98]	25.56	<.001
Direction (cent-down)	-0.14	[-0.33;0.04]	-1.51	.131	0.04	[-0.10;0.18]	0.50	.618	-0.09	[-0.24;0.05]	-1.24	.214	0.04	[-0.17;0.25]	0.35	0.726
Direction (up-cent)	0.00	[-0.19;0.19]	0.00	.996	-0.12	[-0.26;0.02]	-1.64	.102	-0.12	[-0.26;0.03]	-1.58	.114	-0.13	[-0.34;0.09]	-1.15	0.251
Sentence Location	0.20	[0.08;0.32]	3.31	.002	0.22	[0.11;0.33]	4.03	<.001	0.25	[0.09;0.41]	3.14	.003	0.49	[0.23;0.75]	3.67	<.001

Note. Analyses were based on all trials which fell in a certain range. The upper boundary of the low similarity range corresponds to the point value in the 2nd column of Figure 7, the lower boundary of the low-medium range corresponds to the point value in the 2nd column of Figure 7 and the upper boundary of the low-medium range corresponds to the point value in the 3rd column of Figure 7 etc.

Furthermore, we additionally explored in how far the observed effects of sentence spatial location on lexical choices hinged on predictability. To this end, we computed cloze values for the stimuli which ranged from 0.07 - 0.43 per sentence proving that none of the sentence endings was highly predictable. The absolute number of produced words per stimulus sentence ranged from 17 - 42 different words, see Supplement S3 for information on the predictability of words for each sentence from our stimulus material.

We then ran an additional post-hoc analysis with only those sentences included where less than 36 different nouns had been produced as an answer (in total 72 different nouns could have been produced potentially) which are the more highly predictable stimuli in the stimulus set. This reduced dataset with 40 out of 59 stimulus sentences yielded very similar results to the main analysis with a significant interaction between spatial location of sentences and semantic similarity ($\beta_{\textit{high-pred}} = 0.57$, $t = 3.28$, $p = .001$).

On the contrary, when looking at those cases where participants had produced the most diverse answers, i. e. 36 or more different nouns per sentence (19 out of 59 stimulus sentences), there was no interaction between spatial locations and semantic similarity ($\beta_{\textit{low-pred}} = 0.31$, $t = 1.43$, $p = .153$), while the main effect for spatial locations of stimulus sentences on spatial locations of produced nouns was significantly evident in both subsets ($\beta_{\textit{high-pred}} = 0.24$, $t = 3.94$, $p < .001$ and $\beta_{\textit{low-pred}} = 0.30$, $t = 6.05$, $p < .001$). Thus, the effect of spatial locations of the stimulus material on spatial locations of the produced nouns persists when predictability is minimized.

Discussion

We replicated the main effect of sentence spatial locations, finding that the referents of produced nouns were higher up/lower in the world when sentences described situations higher up/lower in the world, as indicated by prior ratings of the nouns in these written sentences. For example, when choosing suitable sentence endings for the sentence <*You are at the vantage* *point and you see ...>* participants chose words such as *sky, mountains, skyscraper* or *Ferris wheel* while they completed the sentence *<You are at the canal and you see...>* with words like *ant, stones* or *litter*. By showing that spatial meaning traces influence the choice of words in an open language production task, we were able to demonstrate that experientially driven meaning aspects in the spatial domain have an impact on lexical selection during language production. Furthermore, there was an interaction between semantic similarity and sentence spatial location, indicating that the effect of sentence spatial location on the spatial properties of the produced noun was higher when the sentence noun and the produced noun were semantically related. In contrast to Experiment 1, there was a significant effect of the spatial manipulation on the spatial properties of the produced nouns. The different results in Experiment 1 and 2 concerning the influence of semantic similarity and the impact of presentation direction on the spatial properties of the produced nouns are discussed in the next section.

General Discussion

In two experiments we investigated the role of experientially grounded meaning in language production. We manipulated the meaning dimension of LOCATION IN SPACE, in two complementary context conditions in one of two ways: (1) physically and isolated from the meaning of verbal contexts as a simulated ascending or descending movement or (2) embedded within verbal contexts. Specifically, participants read sentence fragments like *<You hike through the forest and you see...>* and completed them with a suitable noun of their choice. Starting from the center of the screen, the words were presented in a simulated up- or downward movement, i. e. a physical vertical visual manipulation. Additionally, spatial cues were conveyed via the meaning of the sentences, i. e. verbally referring to situations in different spatial locations like *<You walk to the field and you see...>* or *<You are on the balcony and you see...>*. We tested whether the physically or verbally transmitted spatial experiential manipulations affect our lexical choices.

Experiential traces embedded in meaningful contexts, but not physical cues, lead to experientially grounded lexical selection

Contrary to the hypothesis that visual motion affects lexical selection, the physical simulation of visual motion did not influence which words participants chose in Experiment 1 and the result was replicated in Experiment 2 when considering the set of sentence nouns which had already been used in Experiment 1. There were no differences in spatial characteristics for nouns produced after descending vs. ascending sentences. The unexpected effect of spatial characteristics of produced nouns being higher after centrally presented sentences than after descending and ascending ones in Experiment 2 seems to be an artifact of this additional condition. The central condition differed from ascending and descending movement manipulations because sentences were presented statically without movement simulations involved. Furthermore, the difference between ascending and descending presentation direction for the whole set of stimuli in Experiment 2 seems to hinge on some of the newly introduced sentence nouns, as it was not existent for the set of items in Experiment 1. Additionally, these effects were small compared to the effects of sentence spatial locations on produced nouns (to be discussed in detail below). Therefore, the unexpected effects for Experiment 2 are not reliably observed and may have been caused by the additional central sentence presentation as well as variations in stimulus material. Potentially, in future studies, a sentence display where the control condition is displayed with a slight shift movement to the right – and not statically – might help to clarify this issue. With regards to stimulus material, Experiment 1 and 2 differed as follows: stimulus sentence were presented with article in Experiment 1, but without article in Experiment 2, and the stimulus set in Experiment 2 was more generic with verbs not describing manner of motion as some verbs in Experiment 1 did (in Experiment 2 five different verbs had been used: *stand, walk, enter,* *go*, and *be*, whereas in Experiment 1 thirty different verbs had been used, among which verbs like *balance* or *paddle*). Furthermore, the stimulus set was reduced to 60 stimulus sentences in Experiment 2 (Experiment 1: 90 stimulus sentences). We do not have an assumption why these differences may have led to the unexpected effect. The comparison of effects between Experiment 1 and 2 with identical situations described in the stimulus sentences (see Figure 6) suggests that not only the central presentation but also some differences in the stimulus material between Experiment 1 and 2 may have contributed to different outcomes. However, this may not generalize and should rather be interpreted as no support for an effect of visual spatial manipulations on lexical selection. New data from an experiment in which we investigated whether body posture changes influence lexical choices and were we used most of the stimuli from Experiment 2 further support this interpretation. In this study, there was a significant difference between nouns produced after upward head movements compared to downward head movements, in line with the hypotheses (Vogt, Ganter, Kaup, & Abdel Rahman, 2022).

The absence of an effect of physical cues speaks against a high susceptibility for experientially grounded aspects on lexical access during language production. This stands in contrast to empirical evidence for experientially grounded language comprehension where influences of visual cues on processing of sentences, nouns and verbs have been reported in different paradigms (Dudschig et al., 2013; Kaschak et al., 2005; Meteyard et al., 2008). One possible explanation for the lack of comparable effects in language production is that the physical manipulation does not transport sufficient meaning to affect the lexical-semantic construction of verbal messages. Analogously, it has been shown that physical spatial cues alone are not sufficient to facilitate an anagram solving task, whereas the combination of spatial and situational cues are (Berndt et al., 2018). Presumably, a higher task relevance of the physical manipulation leading to more effortful linguistic processing (Louwerse, Hutchinson, Tillman, & Recchia, 2015) and / or a stronger bodily involvement (i. e. by changing the body position as in the study by Lachmair, Ruiz Fernández, et al., 2016) may yield an effect of spatial manipulations on lexical selection.

We explored this question in a follow-up study in which participants listened to similar sentence fragments while producing an upward or downward head movement with eyes being closed and producing suitable sentence endings with heads up vs. down. We replicated the effects of sentence spatial properties on the spatial properties of produced nouns which are the focus of this paper. Additionally, we found an effect of head movement on the locations of produced nouns in this study which we interpret as evidence for the position that a substantial amount of experiential reactivation is needed to have an influence on lexical access (Vogt, Ganter, et al., 2022).

An additional factor which might have contributed to the absence of an effect of visual sentence movement is a lack of variability in spatial location of produced words as most words were rated as rather downward related. Given the data from the head movement study where we also observed that produced words tended to be more downward than upward we do not consider this lack of variability the best explanation for the absence of the expected effect.

In contrast to the purely physical visual stimulation, the spatial context manipulation conveyed by the sentences carried more meaning. Indeed, the produced words were influenced by the spatial characteristics of the presented sentence fragments. For example, after reading a sentence like <*You are at the sea and you see...*> participants were more likely to say *a shell* than *a gull.* Crucially, both shells and gulls can be found at the sea. Furthermore, both words get assigned a comparable semantic similarity when using distributional measures of semantics as we did in our study (sea–shell: 0.40; sea–gull: 0.41; on a scale from 0 = *no similarity* to 1 = *synonyms*). However, words additionally sharing the spatial location with the situation described by the sentences were more likely to be selected.

We take this as evidence that experiential knowledge not only affects the way word meaning is represented, but also that it is activated during lexical-semantic planning stages, thereby influencing which words we choose. It has been shown by Ostarek and Vigliocco (2017) that identification of pictorial stimuli was facilitated when presented 250 ms after reading words which belong to the same event (e.g., reading *sky* and seeing a depicted cloud) when the image was presented in the same vertical location where it is typically seen which demonstrates the importance of events during perceptual simulation. Therefore, we deem it likely that in our experiments participants simulate the scenes described by the sentences and that these simulations modulate conceptual and lexical processing. According to situation model theory, specifically the event-indexing model (Zwaan & Radvansky, 1998), space is an important meaning dimension when it comes to integrating different pieces of information given in the linguistic input. From this perspective, it is not surprising that participants produced words that share the spatial properties of the simulations they created when reading the previous linguistic input. In general, it seems that situation model theory (e.g., Zwaan, 2016) fits well with our results, assuming a division of labor between more symbolic and more grounded representations in discourse and thus providing a good explanation for the combined effects of semantic similarity and the more experientially grounded meaning dimension observed in the present study.

To summarize, while many other factors – regarding the selected content words – may play a role during the lexical selection process, we want to highlight that experientially grounded meaning seems to be one important factor in language production. Also, we would like to point out that our results fit well with situation model theory.

The relation between experientially grounded meaning and predictability

It might be argued that the produced words are all more or less predictable given the sentential context in the sense that most of them would probably not lead to processing difficulties when presented in a comprehension task and that it is therefore important to clarify whether the observed effect is carried by spatial location specifically or more generally by predictability. However, even when only examining the stimuli where participants showed most variability in answering, i. e. at least on average every second participant produced a different word, we still obtained the main effect that spatial locations of the stimulus material predicted the spatial properties of the produced nouns.

We think that the notion of predictability with regards to lexical selection is empirically underspecified so far and that more research should be done to investigate which factors contribute to words being predicted in a given context. Our data show that experientially grounded meaning facets might be among those factors. Additionally, statistical distributional properties of language might be important for predictability.

The relation between experientially grounded meaning and traditional semantic measures

The experience-related manipulation of space employed here is embedded in meaningful contexts, but at the same time distinct from semantic context measures known to affect lexical-semantic processing during language production, such as semantic features, associations, thematic relations or categories (Abdel Rahman & Melinger, 2019; McRae et al., 2005). To further examine the influence of the semantic contents of the presented sentences on the produced nouns we included a distributional semantic similarity measure as a covariate in our analysis. We used cosine values computed from text corpora as they are an established measure in the field of semantics and therefore they provided a pragmatic way to yield similarity values for the large data set at hand. As the estimates of semantic similarity are based on huge language corpora they pick up on the statistical linguistic regularities which we encounter in our daily life and are therefore very strong tools in modeling our linguistic behavior. Even though distributional measures of semantics have hardly been incorporated into research on semantic processing in language production (Vinson et al., 2014) they are in our view perfectly suited for quantifying the semantic relationship between the nouns in the sentence fragments and the produced nouns. In Experiment 2 we found a more pronounced effect of spatial characteristics of the presented sentence on the spatial properties of the produced nouns for semantically related relative to unrelated pairs. More precisely, the closer the produced noun was to the content in the visually presented sentence, the stronger the impact of spatial location of the presented sentence fragment on the spatial location of the produced noun. Note that this interaction between similarity and sentence spatial location was only apparent in Experiment 2. This suggests that the presented article in the first experiment made it more difficult for the semantically most associated nouns to be produced. Indeed, the mean cosine value across all trials was lower in Experiment 1 (mean cosine value: 0.21) than in Experiment 2 (mean cosine value: 0.32), that is, produced nouns were overall less semantically related to the content of sentences in Experiment 1 than in Experiment 2. Thus, the design of Experiment 1, in which a determiner restricted the range of possible nouns, made it less likely that the produced nouns were chosen merely because their semantic association to the sentence context was strongest.

Crucially, however, we found that the spatial location values of the sentence nouns predicted the spatial properties of the produced nouns in Experiment 1. Moreover, in Experiment 2, the effect of sentence spatial location on spatial characteristics of the produced nouns was still existent in cases of minimal and high similarity relation. That is, even among the most loosely related cases, nouns were chosen which shared the spatial dimension on top. Therefore, the semantic similarity measure used here cannot entirely explain the relationship between the spatial characteristics of the sentence material and the produced nouns. Rather, similarity seems to work as a moderator, influencing the strength of the impact of spatial properties of the stimuli on the dependent variable. We conclude that meaning aspects as captured by the similarity measure and experientially grounded sensory meaning are closely entangled, but distinguishable, in line with theoretical accounts (Louwerse, 2018; Vigliocco et al., 2009).

Similar results have recently been obtained by Banks et al. (2021). Using a category production task, they were able to predict performance when taking both shared sensorimotor knowledge and linguistic proximity based on distributional knowledge into account. This goes in line with our interpretation that experiential and linguistic associations are both important, contributing separately to the responses we found.

Conclusions

In the present study, we show that lexical-semantic processes during language production are not influenced by physical spatial cues isolated from meaning. Instead, we provide evidence that lexical choices are influenced by experientially grounded sensory meaning of space – as conveyed by the verbal context – and that these choices are modulated by distributional properties of the linguistic context. This is in line with current hybrid theories of semantic memory, which treat sensorimotor aspects and usage-based distributional aspects of language as separate – but interacting – types of meaning (Carota et al., 2021; Davis & Yee, 2021; Vigliocco et al., 2009).

We propose that message planning for speaking does not only involve classic semantic meaning relations as categorical or associative links but may also include other aspects of meaning grounded in sensory, motor or bodily experiences. Future research should study whether the impact of the meaning dimension of LOCATION IN SPACE is captured best as a reactivation of sensorimotor experiences, and thus constitutes evidence for experiential grounding in language production, or whether spatial locations are activated as part of propositional and amodal semantic features (Meteyard et al., 2012). However, based on the evidence reviewed in the Introduction and evidence for activation of spatial-oculomotor regions in the brain during the processing of implicitly spatial nouns (Ostarek, 2018) we deem the meaning aspect of LOCATIONS IN SPACE a strong candidate for experientially grounded meaning.

In the experimental task employed here we investigated which words are chosen during lexical-semantic processing. Traditionally, most studies dealing with lexical access in language production manipulate the context of an utterance, whereas the to-be-produced word is predetermined by the experimental setup (various picture naming tasks, e.g., cyclic blocking, picture word interference, continuous naming). Production tasks with a focus on semantics rarely allow for free lexical choices even though recent studies have moved in this direction (e.g., Fjaellingsdal et al., 2020). Here, participants were not entirely unrestricted in their lexical choices, but could freely select their utterances within non-constraining contexts, allowing us to investigate which factors shape the content of a produced message, rather than the duration of lexical processing. As is typical in everyday language use, our task also encompassed an interplay of comprehension and production (Indefrey & Levelt, 2004; Pickering & Garrod, 2013). Therefore, our experiments provide an important step towards a more complete understanding of one of the crucial elements of language production and we hope to spark interesting discussions and studies which will shed more light on the factors which contribute to answering the question why we choose certain words in order to express an intended meaning.

Acknowledgements

This work was supported by the Berlin School of Mind and Brain and a doctoral fellowship by the Studienstiftung des deutschen Volkes (German National Academic Foundation) awarded to Anne Vogt. We thank Eduard Berndt for providing the data set of spatial ratings with 250 German nouns used for the control words. We are grateful to Fritz Günther for help with implementing and discussing the analyses which are based on distributional similarity measures. Furthermore, we thank Tabea Villinger for supporting stimulus preparation and data collection of Experiment 2 and Guido Kiecker for technical support. All authors developed the study design, discussed the results and wrote the manuscript. Anne Vogt performed data collection and analyses.

Supplement

Table S1

Experiment 1. LMM statistics for direction, sentence spatial location and cosine values serving as similarity measure computed with the semantic space de_wiki. Random effects for subjects were not included due to convergence errors.

Variable	*b*	*CI*	*t*	*p*
Intercept	3.57	[3.49; 3.66]	83.05	<.001
Direction (ascending-descending)	0.00	[-0.09; 0.09]	-0.04	.968
Sentence Spatial Location	0.20	[0.11; 0.28]	4.58	<.001
Semantic Similarity	-0.65	[-1.06; -0.25]	-3.15	.002
Sentence Spatial Location x Semantic Similarity	0.20	[-0.21; 0.62]	0.97	.335

Note. The effect of spatial similarity cannot be interpreted meaningfully as the predicted values were spatial characteristics of the produced nouns. See Experiment 2, for discussion of a different impact of similarity on the predicted values.

Table S2

Experiment 2. LMM statistics for direction, sentence spatial location and cosine values serving as similarity measure computed with the semantic space de_wiki

Variable	*b*	*CI*	*t*	*p*
Intercept	3.72	[3.61; 3.83]	66.80	<.001
Direction (central-descending)	-0.06	[-0.15; 0.03]	-1.36	.175
Direction (ascending-central)	-0.09	[-0.18; 0.00]	-2.04	.041
Sentence Spatial Location	0.30	[0.21; 0.39]	6.40	<.001
Semantic Similarity	-0.23	[-0.55; 0.09]	-1.39	0.165
Sentence Spatial Location x Semantic Similarity	0.90	[0.63; 1.18]	6.39	<.001

Table S3

Predictability Experiment 2. Table contains the most frequently produced noun for each stimulus sentence together with its cloze value demonstrating that predictability as indicated by cloze values is not a likely source of the effects we found with cloze values being generally low. The rightmost column contains the total number of different concepts produced for each stimulus sentence.

stimulus number	noun in stimulus sentence	most frequently produced sentence ending	absolute count of most frequently produced noun	cloze value (relative proportion of most frequently produced noun)	number of different concepts
1	Meer	ein Strand	9	0.125	34
2	Pool	eine Luftmatratze	9	0.125	34
		Wasser	9	0.125	
3	Kanal	ein Boot	14	0.194	37
4	Fluss	ein Fisch	9	0.125	37
5	See	ein Boot	8	0.111	40
6	Teich	ein Frosch	14	0.194	29
7	Bach	ein Fisch	21	0.292	28
8	Feld	ein Hase	8	0.111	47
9	Ufer	ein Boot	8	0.111	42
10	Strand	Sand	8	0.111	30
11	Wiese	eine Blume	15	0.208	38
12	Straße	ein Auto	23	0.319	28
13	Kreuzung	eine Ampel	26	0.361	17
14	Lichtung	ein Reh	16	0.222	35
16	Garten	eine Blume	15	0.208	45

17	Zelt	ein Schlafsack	8	0.111	36
18	Sofa	ein Kissen	15	0.208	22
19	Freibad	ein Sprungbrett	9	0.125	31
20	Haltestelle	ein Bus	31	0.431	18
21	Weide	eine Kuh	21	0.292	21
22	Baustelle	ein Bagger	14	0.194	30
23	Bahnhof	ein Zug	31	0.431	26
24	Zug	ein Gleis	5	0.069	26
25	Garage	ein Auto	29	0.403	23
26	Eingang	eine Tür	22	0.306	22
27	Terrasse	die Sonne	5	0.069	40
		ein Stuhl	5	0.069	
		ein Vogel	5	0.069	
		eine Blume	5	0.069	
28	Innenhof	ein Fahrrad	8	0.111	41
29	Park	ein Hund	15	0.208	32
30	Küche	ein Kühlschrank	10	0.139	38
31	Bus	eine Haltestelle	7	0.097	33
32	Veranda	ein Garten	7	0.097	36
33	Ofen	ein Feuer	10	0.139	31
		ein Kuchen	10	0.139	
34	Badezimmer	ein Spiegel	12	0.167	18
35	Straßenbahn	ein Bus	5	0.069	27
36	Café	ein Tisch	13	0.181	20
37	Museum	ein Gemälde	17	0.236	25

38	Zaun	ein Hund	5	0.069	41
		ein Vogel	5	0.069	
39	Schuppen	eine Schaufel	19	0.264	30
40	Turnhalle	ein Ball	14	0.194	29
41	Schule	eine Schultafel	17	0.236	28
42	Haustür	ein Schlüssel	9	0.125	27
43	Dschungel	ein Affe	14	0.194	22
44	Schaufenster	eine Schaufensterpuppe	29	0.403	20
45	Tisch	ein Teller	12	0.167	30
46	Wald	ein Reh	12	0.167	35
47	Haus	eine Tür	13	0.181	38
48	Scheune	ein Pferd	13	0.181	26
49	Garderobe	eine Jacke	25	0.347	20
50	Kaufhaus	eine Kasse	6	0.083	41
51	Brücke	Wasser	10	0.139	32
52	Stadion	ein Tor	8	0.111	21
53	Baum	ein Vogel	12	0.167	29
54	Fenster	ein Vogel	7	0.097	35
55	Straßenlaterne	Licht	14	0.194	36
56	Balkon	eine Blume	10	0.139	39
57	Leuchtturm	das Meer	21	0.292	26
58	Gebirge	Berge	10	0.139	34
59	Baumhaus	ein Vogel	5	0.069	43
60	Aussichtspunkt	der Horizont	5	0.069	37

3.2 Manuscript 2: Embodied language production: sensorimotor activations and interoceptive sensibility influence which words we choose when speaking

This manuscript has been submitted for publications and is currently under peer review. It is available as a preprint:

Vogt, A., Ganter, I., Kaup, B., & Abdel Rahman, R. (2022). Embodied language production: Sensorimotor activations and interoceptive sensibility influence which words we choose when speaking *PsyArXiV*. https://psyarxiv.com/3zgrc

This is the author's version of the manuscript. Small formatting adjustments were done for inclusion in this dissertation. Stimuli, data and analysis scripts can be found in an online repository: https://osf.io/h53ud/?view_only=a8c935748bbe4996a3a0f2b5d64099e2
References are listed at the end of the dissertation.

Abstract

We know little about the factors influencing which words we choose during lexical selection. In two experiments we investigated whether (re-)activations of experiential traces of space have an impact on language production. Participants performed up- and downward head movements while listening to sentence fragments describing situations (e.g., <*You are at the beach and you see...*>). When reaching the upward/downward head position they completed the sentence with a freely chosen noun. A different group of participants rated the spatial location of the produced word's referents. We found that the head movements influenced participant's lexical choices. After upward movements the produced words were rated to be located higher up in space compared to downward movements. Furthermore, higher scores in interoceptive sensibility as measured using the *attention regulation* scale from the Multidimensional Assessment of Interoceptive Awareness questionnaire (Mehling et al., 2018) lead to an increased effect of head movement on the spatial properties of the produced nouns. We conclude that sensorimotor activations are among the meaning facets that guide which words we chose when speaking. The tendency to verbally express embodied meaning is enhanced with higher levels of interoceptive sensibility, suggesting that interoception may be a key to understand interindividual differences in how we express our experiences and feelings when we speak.

Keywords: language production, lexical selection, interindividual differences, embodied cognition, interoceptive sensibility, language-space associations

Language is a core human faculty and we strongly rely on language to communicate our thoughts, experiences and feelings to others. Despite the importance of language for each of us and a long history of reasoning and research on language and meaning reaching back to ancient Greece (Kiefer & Pulvermüller, 2012), there remain many open questions about the relation between words, meaning and lexical processing. According to the theory of Embodied Cognition sensory and bodily experiences become closely entangled with the words referring to them and they are reactivated whenever we encounter the respective words or concepts (Barsalou, 2008; Zwaan & Madden, 2005).

Evidence in favor of embodied or experientially grounded meaning processing, demonstrating that the activation of experiential traces plays an important role in language processing, stems mainly from language comprehension research (Binder & Desai, 2011; Kaup et al., 2016; Pulvermüller, 2018). For example, in the brain, motor areas and adjacent regions that are activated in the context of actual movements of body parts are also activated when passively reading action words referring to those body parts, e.g., *to lick* – face, *to kick* – leg, *to pick* – arm (Hauk et al., 2004). Similar and rapid effects have been observed for action and sensory conceptual features in unconsciously perceived words using EEG (e.g., Trumpp, Traub, Pulvermüller, & Kiefer, 2014).

Another well-investigated, experientially grounded meaning facet are language-space associations, i. e. information on where in the world events typically occur or where certain objects can typically be found. Experiential traces of space are re-activated when processing the respective words (Bergen et al., 2007; Dudschig et al., 2012; Estes et al., 2008; Lachmair et al., 2011; Ostarek, Ishag, et al., 2018; Öttl et al., 2017; Thornton et al., 2013). In a study on body position and retrieval of previously memorized words, Lachmair and colleagues (2016) showed a congruency effect for word recall, i. e. participants remembered more words in accordance with the spatial properties of the words when they were in a certain body position (head-down vs. upright).

While the activation of experiential traces is well-investigated in language comprehension, only few studies have thus far investigated whether meaning grounded in bodily and sensory experiences is relevant for lexical selection during language production.

The role of semantic relations in lexical selection

Language production consists of a series of interactive processes starting with the speaker's intent to convey a message which is then transferred into lexical representations (lemmas) which are transformed into sequences of spoken phonemes or written graphemes (e.g., Bonin & Fayol, 2000; Dell, 1986; Levelt, 1989; Levelt et al., 1999; Oppenheim et al., 2010; Roelofs, 2018; Torrance et al., 2018). Using picture naming tasks it has been shown that lexical representations are activated together with semantically related representations and that the target lemma is selected from this cohort of co-activated lexical alternatives (Dell, 1986; Levelt, 1989; Mahon et al., 2007). The role of semantic relations in language production has most frequently been investigated by presenting pictures in the context of categorically related items. For example, in the picture word interference paradigm naming the picture of a dog in the presence of a categorically related distractor word, e.g., *cat* is slower than in the presence of an unrelated distractor (Bürki et al., 2020). Other meaning relations like associations (e.g., dog – leash), part-whole relations (e.g., shoe – lace), or thematic links (e.g., wedding – flowers) have been investigated much less frequently (e.g., Abdel Rahman & Melinger, 2009, 2019; Alario et al., 2000; Damian & Spalek, 2014; de Zubicaray et al., 2013; La Heij et al., 1990; Lin et al., 2021). Importantly, semantic relations may also include meaning facets which are based on sensory experiences like body movements, sounds, colors or shapes – all of which have been found to play a role in language production (de Zubicaray et al., 2017, 2018; Hirschfeld & Zwitserlood, 2012; Mädebach et al., 2017; Matheson et al., 2014; Mulatti et al., 2014; Redmann et al., 2014). In a recent study by Banks et al. (2021), participants were asked to name as many category members as possible for a given semantic category. It was found that both the order and the frequency of produced words can be predicted by both the experience-based semantic similarity between the words and the category (e.g., the word *cat* showed higher sensorimotor similarity to the category <animal> than the word *butterfly* and it showed higher semantic similarity according to a measure of linguistic distributional aspects of meaning, e.g., as indexed by cooccurrence metrics). This implies that we make use of the experiential and linguistic contexts in which words occur when we speak and process language (Davis & Yee, 2021).

Still, it remains a hitherto unanswered question whether embodiment in the form of (re-)activations of experientially grounded meaning is relevant for language production (Vinson et al., 2014; Vogt, Kaup, & Abdel Rahman, 2022). Here we test whether embodied meaning directly affects which words we chose when we plan to speak.

Individual differences and their impact on embodied meaning processing

Interindividual differences may be among the factors contributing to some of the heterogeneous findings in the field of embodied language processing (Barsalou, 2020; Casasanto & Henetz, 2012; Keehner & Fischer, 2012; Meteyard et al., 2012; Ostarek & Huettig, 2019; Pitt & Casasanto, 2019). For example, Yee et al. (2013) found that naming of objects may be hindered when participants are simultaneously occupied with a manual movement incompatible with the movements used to interact with the to-be-named objects, with effects being even stronger the more experience participants had with interacting with those objects. Converging evidence also comes from research with different populations. Participants with high vs. low sensorimotor skills showed differential processing of action-related texts (Beilock, Lyons, Mattarella-Micke, Nusbaum, & Small, 2008; Holt & Beilock, 2006). Similarly, participants with Parkinson's disease, a neurodegenerative disease leading to decreased activation of brain regions required for motor control, were compared to a matched healthy control group and showed impaired naming for action-related words (Cotelli et al., 2007; Rodríguez-Ferreiro, Menéndez, Ribacoba, & Cuetos, 2009).

Individuals may not only differ with regard to their experiences but also in the degree to which certain aspects of the environment are taken into account during cognitive processing and in the degree to which they are sensitive towards their own body. As a possible consequence, experientially grounded cognition based on bodily sensations may be moderated by interindividual differences in interoception, i. e. our sensitivity to processes originating inside or at least concerning our bodily states (Häfner, 2013). While it has been shown that the perception and processing of internal body signals is intricately linked with embodied cognition (Barrett & Simmons, 2015; Buldeo, 2015; Herbert & Pollatos, 2012; Villani et al., 2021) there is – to the best of our knowledge – no empirical evidence for effects on language production due to interindividual differences in interoception.

Here we test whether interindividual differences in interoceptive sensibility (Mehling et al., 2012) relate to possible embodiment effects in language production. Specifically, we investigate moderating influences of body posture on lexical selection in a free language production task.

The present study

This study was designed to test whether sensorimotor activations, here in the form of upward or downward head movements, influence which words we choose when speaking. In addition, we aimed at finding out whether interindividual differences in interoceptive sensibility, the general tendency to perceive body processes, moderate the impact of the head posture on lexical selection. We adopted a paradigm by Vogt, Kaup, and Abdel Rahman (2022) in which participants completed non-constraining sentence fragments like <*You are in the forest and you see...*> or <*You are in the mountains and you see...*> with suitable nouns of their choice. In the original paradigm, sentences were visually displayed word by word in a rapid serial presentation mode with each word being presented slightly below or above the previous word, inducing a visual downward or upward appearance of the sentences on the screen. This manipulation of vertical space had no influence on the location of words used to finish the sentence fragments, suggesting that a physical manipulation of spatial location that has been shown to affect language comprehension (Dudschig et al., 2013; Kaschak et al., 2005; Meteyard et al., 2008) does not affect language production. However, the location in space of the sentence fragments (e.g., *sea* – relatively low vs. *mountains* – relatively high) predicted the spatial location of the words used to complete the sentences: participants were more likely to complete a sentence like <*You are at the sea and you see...*> with *a shell* than with *a gull*, even though both words are similarly semantically related to the sentence noun *sea*. Indeed, distributional measures of semantic similarity revealed that there was a contribution of the reactivation of experiential traces of space on lexical choices triggered by the contents of sentence fragments that was independent of mere semantic associations.

Here we used the same sentences, but instead of a vertical visual manipulation of the sentence display on the screen, participants listened to the sentence fragments while making an upward or downward movement with their head, completing the sentence upon reaching the head-up or head-down position. This manipulation in the form of active head movements that are meaningfully related to the location in space should increase the impact of embodied meaning on lexical choices. The location of the produced words should be predicted by the upward or downward movement of the head.

Additionally, we administered self-report questionnaires to assess interoceptive sensibility, a dimension of interoception related to the tendency to attend to and notice body sensations (Garfinkel et al., 2015). Furthermore, we expected that the relation between head movement and location of the produced sentence ending should be enhanced with higher individual tendencies to attend to the bodily states as indicated by interoceptive sensibility scores. Finally, as in the previous study by Vogt and colleagues (2022) we also included distributional semantic similarity measures between the nouns in the sentences and the produced nouns, expecting to replicate the observation that the spatial properties of the words in the sentence fragments influence the spatial properties of the produced words independent of semantic associations.

Both experiments were preregistered (Experiment 1: https://aspredicted.org/X19_3BB and Experiment 2: https://aspredicted.org/3KR_ZY2).

Experiment 1

Methods

Participants. In total, 53 native German speakers were recruited personally and using the institute's participant pool Psychologischer Experimental Server Adlershof (PESA). Data of nine participants were excluded prior to analysis based on our preregistered exclusion criteria because the participant correctly inferred the goal of the study (n = 1), non-compliance with the experimental procedures (n = 1), technical issues resulting in the loss of audio files (n = 1), limited German proficiency despite reporting being a native speaker (n = 1), or excessive trial loss (> 20 %) due to a high number of missing or invalid answers (n = 5). Data of 44 participants (32 female, 18 - 33 years, $M_{\text{age}} = 24.68$, $SD_{age} = 3.59$) entered the analysis. Sample size was planned on the basis of a behavioral pilot study with eight participants using the same study design. Based on analysis of these pilot data with linear mixed effect models an effect of $\beta = 0.12$ for the movement manipulation (head up vs. down) on spatial properties of the produced words was used to run a power analysis. Using the package simr (Green & MacLeod, 2016) 1000 simulations revealed that 44 participants would be needed to achieve 82.9 % power ($95\%CI[80.42, 85.18]$) to secure the fixed effect of head movements on lexical choices. Participants provided informed consent to their participation in the study. The study was conducted based on the principles expressed in the Declaration of Helsinki and was approved by the local Ethics Committee. Participants received course credit or a monetary compensation.

Materials.

Sentence Stimuli. 42 German sentence fragments (from Vogt et al., 2022) were presented, each describing a setting like *<Du bist am Meer und siehst...>* (English: *<You are at the sea and you see...>*). Sentences consisted of six words, starting with the personal pronoun *du* (English: *you*) followed by a verb, a place description and ending with the phrase *und siehst* (English: *and you see*), indicating that participants should complete the sentence with a suitable noun phrase (noun plus determiner). The nouns at the fourth position of the stimulus sentences were rated beforehand according to their spatial location (9-point Likert scale ranging from 'down' over 'central' to 'up', see Vogt, Kaup, & Abdel Rahman, 2022) to ensure that the sentence fragments described situations spanning a wide range of spatial locations on the vertical axis. All sentences with their respective spatial location ratings can be found in Appendix A. Sentences were pre-recorded by a female speaker with recording durations between 2.3 to 2.8 sec per sentence. Six additional sentences were used in practice trials.

Interoception Questionnaire. The two subscales *noticing* and *attention regulation* of the German version of the Multidimensional Assessment of Interoceptive Awareness, MAIA (Mehling et al., 2018, 2012) with a total of eleven items were used in order to assess participants' awareness of uncomfortable, comfortable and neutral body sensations as well as the ability to sustain and control attention to body sensations. We deemed these facets most interesting with regards to interindividual differences of interoceptive sensibility in relation to a paradigm assessing the susceptibility towards experimentally manipulated head positions and their impact on lexical choice. Participants indicated how often each of the questionnaire statements (e.g., *When I am tense I notice where the tension is located in my body.*) applies to them in daily life on a 6-point Likert scale ranging from 'never' (0) to 'always' (5).

Procedure. The sentence completion task was programmed using jspsych (de Leeuw, 2015) and was hosted online using JATOS (Lange, Kühn, & Filevich, 2015). While participants completed the sentence fragments without time pressure, they were monitored by the experimenter in a parallel video session via Zoom during instruction and word production. The instruction and word production phase lasted about 25 - 35 minutes. Subsequently, participants completed the interoception assessment which took 5 - 10 minutes.

Participants were expected to perform the experiment alone in a quiet environment while they were free to choose a sitting or standing position. During the instruction phase and in the sentence completion task, they were monitored by the experimenter via video call in order to control the correct execution of head movements. Interaction with the experimenter was only possible during the instruction phase and during breaks but not during or between trials. In the sentence completion task, participants listened to auditorily presented sentence fragments while performing an upward or downward head movement, indicated by a high or low tone of 200 ms duration, respectively (see Figure 1 for depiction of a trial sequence). The tone was presented twice with an onset asynchrony of 1000 ms. Upon hearing the first tone participants closed their eyes, and after the second tone the sentence fragments were presented while participants executed the movement until their head position reached an angle of about 45° in comparison to their torso. Audio recording started upon completion of the sentence fragments for 5 sec. Participants completed the sentences with a noun phrase of their choice when reaching the final head position keeping their eyes closed. Afterwards they returned their head to a straight position and opened their eyes before the next trial started.

Participants were instructed to complete the sentences spontaneously by producing any noun phrase they deemed a suitable ending except for nouns referring to persons (e.g., *a lifesaver, a woman, a kid*). Furthermore, they were asked not to produce complex noun phrases consisting of several words (e.g., including adjectives: *a delicious cake*) and not to repeat words which had been presented in the given sentence or which they had produced in previous trials. In total, 42 sentence fragments were presented and paired with both head movements, resulting in 84 trials. Sentences were paired with one movement in the first part of the experiment and were repeated with the other movement in the second part, assignment of sentence to part was counterbalanced across participants. Within each part, both movements occurred equally often and sentences were presented in random order for each participant.

After every ten trials, participants had the opportunity to take a break. Written instructions and two short videos explaining the task were presented. Participants were asked to practice the required head movement. Afterwards they learned the tone-response assignment (high tone: upward movement, low tone: downward movement) and performed six practice trials. During practice trials the experimenter provided feedback concerning the correct execution of the head movements. After the sentence completion task participants answered the interoception items from the MAIA questionnaire which were displayed in a random order. Then, the purpose of the study was explained.

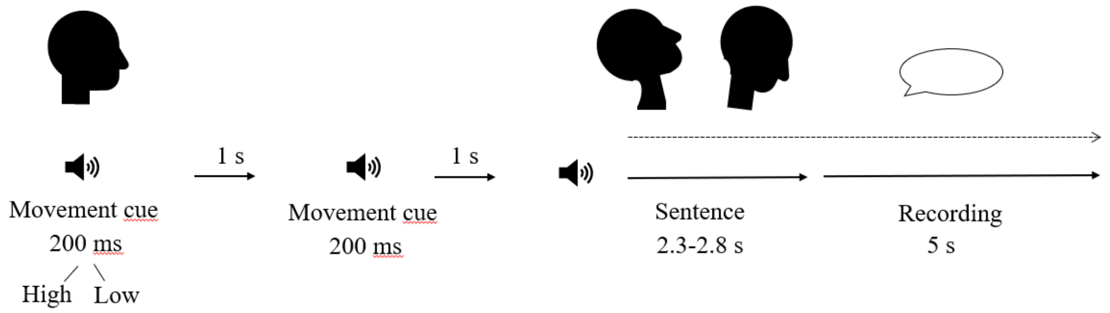

Figure 1. Trial sequence in the word production phase. Participants completed auditorily presented sentences by nouns of their own choice after performing a head movement.

Spatial rating of produced nouns. After the experiment, all words were rated according to their vertical spatial location by an independent group of 35 raters. Data of one rater were not considered for analysis as data loss exceeded 20 %. Additionally, three raters were excluded due to showing low interrater reliability with the control words calculated using the psych package (Revelle, 2018), $ICC < .75$. The final data set included 31 raters (29 female, 18 - 76 years, $M_{age} = 32.32$, $SD_{age} = 18.04$).

The total set of produced words was reduced for the spatial rating using the following criteria: (1) Words which had been produced multiple times by different speakers were included only once in the spatial rating. (2) Ambiguous words were presented with an explanation, which was derived from the content of the presented sentence fragments, e.g., *Blätter (Pflanze)*, English: *leaves (plant)*. An explanation was also given when the produced words might not have been commonly known, e.g., *Bake (Absperrung einer Baustelle)*, English: *beacon (barrier of a construction site)*. (3) Words which were judged to be synonyms or very similar with regard to their referents entered the rating only once, e.g., produced words: *Segelschiff*, English: *sailing ship* and *Segelboot*, English: *sailing boat* > word included in the rating for both instances: *Segelboot*, English: *sailing boat*. A total of 881 nouns remained and were included in the spatial questionnaire. The nouns were randomized and then distributed to nine different questionnaires, each containing 489 or 490 words, thereby ensuring that the questionnaires were representative of the whole set of produced nouns. In addition, 19 control words which had not been uttered in the production phase and which spanned a wide range of vertical positions (e.g., *U-Boot*, English: *submarine*, *Heißluftballon*, English: *hot air balloon*) were included in order to check the quality of the spatial ratings for this subset of items with known spatial properties. Participants rated the spatial location of each word's referent on a vertically displayed 7-point Likert scale ranging from 'down' over 'central' to 'up' with the additional option of not being able to judge the spatial location.

Ratings showed a consistency with previously established spatial location ratings of control words in the range of ICC = .75, 95% CI [.52, .88] to ICC = .97, 95% CI [.94, .99] and agreement between raters was excellent according to the criteria of Fleiss (1986) with $ICC = .81, 95\ \%CI[.72, .89]$, based on single ratings, consistency, and two-way mixed-effect model (see Hallgren, 2012; Koo & Li, 2016; Shrout & Fleiss, 1979, for more details on interrater reliability). Each word was rated by at least 16 different persons. Standard deviations for each word over raters varied from $SD = 0$ for the word *Kleid*, English: *dress*, to $SD = 2.25$ for *Kabel (zur Straßenbahn laufend gesehen)*, English: *cable (seen while walking to the tram)*. Mean ratings of spatial location for each produced word were calculated over all raters and they were merged with the data from the sentence completion study so that for each trial a mean spatial rating serving as an indicator of the spatial location of the produced noun was obtained. The rating was administered using SoSciSurvey.de (Leiner, 2019) and took about 20 min to complete.

Data Analysis

Data preprocessing. A total of 44 participants completed 42 sentences twice (84 trials) resulting in 3696 data points. Missing data or trials where participants did not follow the instructions were removed, leaving 3322 trials. Additionally, one trial was excluded due to experimenter error when setting up the rating. If less than 80 % of the raters judged the spatial location of a word, they were not considered for subsequent analysis, further reducing the amount of data points for analysis to 3265. Additionally, distributional semantic similarity values between the nouns in the auditorily presented sentences and the produced words were included into analyses. Distributional similarities were computed using the dewak100k_cbow semantic space built from the deWaC-corpus through cbow algorithms as implemented in the word2vec model (Mikolov et al., 2013) using the LSAfun package (Günther et al., 2015). Cosine values were computed for 3204 trials as not all produced words were included in the corpus.[6] As cosine values below zero cannot be interpreted meaningfully (Günther et al., 2015), the complete set of trials had to be further reduced by 50 trials. In total, 3154 data points containing distributional semantic similarity estimates entered statistical analyses.

Statistical Analyses. Statistical analyses were performed with R version 4.1.2 (R Core Team, 2020). To investigate whether participants produced more words with referents in the higher sphere when performing an upward head movement and vice versa, words were categorised by their mean spatial ratings into three groups of different spatial locations (low: 1-3, central: 3-5, high: 5-7). A χ^2-Test was used to investigate the difference in the absolute number of high and low words according to an upward or downward head movement. Further, hypotheses were investigated using linear mixed models (LMM) as implemented in the R packages lme4 (Bates,

[6] Some words were slightly changed into words included in the corpus to reduce data loss.

Mächler, et al., 2015) and lmerTest (Kuznetsova et al., 2017). In all analyses, we included the maximal random effects structure supported by our hypotheses which still enabled model convergence and did not lead to overfitting. Therefore, random effects were excluded based on least explained variance if necessary following the procedure suggested by Bates, Kliegl, Vasishth, and Baayen (2015) using the rePCA function. First, correlations between random effects were excluded if the maximal model did not converge. Further, the dimensionality of the variance-covariance matrix was extracted by principal component analyses using the rePCA function of the lme4 package. Variance components explaining none or the least of the cumulative variance were dropped until a significant loss in goodness of fit occurred as indicated by likelihood ratio tests.[7]

We examined the influence of head movement on word choice by using a linear mixed effects model with fixed effects for *movement*, *sentence location* and their interaction. Furthermore, we also added fixed effects for *attention regulation* as well as the interaction of the *attention regulation* score with *movement* and fixed effects for *noticing* as well as the interaction between *noticing* and *movement* and the fixed effects for *movement, sentence location* and their interaction to investigate the influence of differential degrees of interoception. Predictors were entered into the random structure of subject and sentence if the predictor varied within the random effect (e.g., *attention regulation* does not vary within one person). Sliding difference contrast coding was applied to the predictor *movement* (up vs. down) and predictor values for *sentence location* and the interoception scales were centered. The random effect structure of

[7] We deviated from the preregistered analysis plan in three ways.

(1) Random effect simplification was implemented as described above as the automatized reduction of maximal random effect structure using the buildmer package (Voeten, 2021) conflicted with the use of lmerTest, leading to models with convergence errors.

(2) We didn't use a sum score of interoceptive sensibility even though *attention regulation* and *noticing* correlated strongly with r = 0.62 in Experiment 1. However, they correlated with r = .47 in Experiment 2 and therefore, we chose to enter them as separate predictors for consistency in analysing the data from Experiment 1 and 2. Furthermore, the scales of *attention regulation* and *noticing* modulated the impact of the movement manipulation differently with a different polarity for the interaction of *attention regulation* and movement compared to the (non-significant) interaction of *noticing* and movement. Using a sum-score would have shielded these differential effects.

(3) Abstract words were not excluded from the dataset. During preprocessing several words that might have been considered as abstract were excluded for other reasons without specifically categorising abstractness by the experimenter. Two words that could be considered more abstract remained in the dataset (*die Ferne*, English: *the distance*; *die Weite*, English: *the far*), which was considered negligible.

the model was reduced following the above procedure before adding the interaction of semantic similarity and sentence location as a fixed effect.

Results

The absolute number of produced words categorised as high ($n_{moveup} =$ 218, $n_{movedown} =$ 191) or low ($n_{moveup} =$ 549, $n_{movedown} =$ 589) produced per movement condition did not differ significantly, but there was a trend in the expected direction $X^2(1) = 2.88$, p = .089.

Numerically, small differences in spatial mean ratings of produced words depending on whether participants had performed an upward (M = 3.59, SD = 1.23) or downward head movement (M = 3.53, SD = 1.22) occurred. Using a linear mixed effects model (see Table 1 for full model output), a trend for produced words being influenced by the head movement was observed. Additionally, there was a significant interaction between participants' *attention regulation* score and the head movement direction on spatial properties of the produced words. Participants with higher ability to sustain and control attention to body sensations showed a significant influence of head movement on produced words with a linear increase in the location of the produced words according to the respective head movement. There was no effect of subject's *noticing* score in relation to the head movement manipulation. Furthermore, vertical spatial sentence locations significantly influenced the spatial properties of the produced words with increasingly higher vertical positions of the produced words with increasing vertical position of the presented sentences. There was a significant interaction between semantic similarity and sentence location indicating that the effect of sentence spatial location is more pronounced for higher semantic similarity between presented sentence and produced noun.

As preregistered, analyses for the first half of the experiment where participants listened to each sentence for the first time were conducted separately. They revealed a similar pattern with no effect for head movement but a marginally significant positive interaction between head movement and attention regulation as well as a marginally significant negative interaction between head movement and noticing and a significant effect of sentence spatial locations on produced words (see Supplement 1). Furthermore, we conducted analyses without fixed effects for interoception facets. In this analysis the estimates for the fixed effects for movement and spatial location of sentences did not change substantially (see Supplement 2).

Table 1

Linear mixed model predicting spatial location of produced words by taking attention regulation and noticing scores and their interaction with head movement into account.

Model formula

word location ~ 1 + movement + sentence_location + movement x sentence_location + attention_regulation + attention_regulation x movement + noticing + noticing x movement + sentence_location x similarity + (sentence_location + movement x sentence_location || subject) + (1 | stimulus)

Fixed Effects	Estimates [95% CI]	Std. Error	*t*-value	*p*-value
Intercept	3.57 [3.46; 3.67]	0.05	65.76	**<0.001**
Movement	0.06 [-0.01; 0.14]	0.04	1.64	0.101
Sentence Spatial Location	0.29 [0.20; 0.39]	0.05	5.90	**<0.001**
Movement x Sentence Spatial Location	0.03 [-0.06; 0.12]	0.05	0.67	0.505
Attention Regulation	0.02 [-0.05; 0.08]	0.03	0.51	0.611

Movement x Attention Regulation	0.13 [0.01; 0.25]	0.06	2.18	**0.029**
Noticing	-0.03 [-0.1; 0.05]	0.04	-0.76	0.452
Movement x Noticing	-0.08 [-0.22; 0.07]	0.07	-1.05	0.296
Sentence Spatial Location x Semantic Similarity	0.60 [0.33; 0.87]	0.14	4.35	**<0.001**

Note. Number of participants = 44; number of categories = 42; total n = 3154; CI = confidence interval around the estimate; *p*-values are based on Satterthwaite approximation as implemented in lme4 package. Significant p-values of p < .05 are shown in bold.

Discussion

We had expected effects of head movements on word choices. Even though there was a small numerical effect in the expected direction with word's referents being higher up after upward head movements and lower down after downward head movements, this effect did not reach significance. Interestingly, however, the effect of head movement interacted with interoception. With a more pronounced tendency to control and sustain attention to bodily feelings, the effect of head movement on word choices was enhanced. Furthermore, we successfully replicated previous results by Vogt, Kaup, and Abdel Rahman (2022) and demonstrated that the experiential domain of space which is evoked by the situation described in the sentence influences the spatial properties of words which are chosen to complete the presented sentence fragments.

We take this as first tentative evidence that experiential traces of space can be activated in a task where participants have to complete unconstrained sentence fragments while producing a certain head movement and that these reactivations may influence lexical selection. Crucially, however, this effect hinges on interindividual differences in the ability to focus and regulate in how far we attend towards bodily sensations. Before turning to conclusions we aimed to replicate the findings with two changes in the design. The tone-response-assignment used in Experiment 1 (high tone: head up, low tone: head down) might have served as a semantic cue, in which case it was not the body movement per se leading to the observed effect, but the mapping between high / low tone and the upward / downward movement, respectively. Therefore, we repeated the experiment with a reversed tone-head movement-assignment in Experiment 2. Furthermore, we changed the order of sentence completion task and the interoception questionnaire, starting with the questionnaire. Thereby, we intended to enhance the impact of the head movements, assuming that participants would be primed to pay attention to their bodily signals beforehand. By including Experiment 2, we also planned to pool the data from the two experiments, which might stabilize a potentially small main effect of head movement across a counterbalanced assignment of tone to head movement.

Experiment 2

Methods

Experiment 2 was identical to Experiment 1 except for a few changes.

Participants. In total, 70 native German speakers were recruited using the institute's participant pool Psychologischer Experimental Server Adlershof (PESA), aiming for at least the sample size of Experiment 1, and ending up with valid data from 55 participants (40 female, 14 male, 1 unknown, 18 - 34 years, $M_{\text{age}} = 24.84$, $SD_{age} = 4.71$) who had not taken part in Experiment 1. Data of 15 participants were excluded prior to analysis based on our preregistered exclusion criteria because they correctly inferred to goal of the study (n = 8), had mistakenly been taking part in Experiment 1 as well (n = 2), or excessive trial loss (> 20 %) due to a high number of missing or invalid answers (n = 5). Participants provided informed consent to their participation in the study. The study was conducted based on the principles expressed in the

Declaration of Helsinki and was approved by the local Ethics Committee. Participants received course credit or a monetary compensation.

Materials.

Sentence Stimuli. We used the same stimuli as in Experiment 1.

Interoception Questionnaire. The two facets *noticing* and *attention regulation* of the German version of the MAIA questionnaire (Mehling et al., 2018, 2012) were used again to assess participants' interoceptive sensibility. Furthermore, we also explored whether the facet *body listening* from the MAIA questionnaire can be used to investigate participants' susceptibility to head movement manipulations. *Body listening* assesses the tendency to actively listen to the body for insight and it is assessed by three items, e.g., <*I listen to my body to inform me about what to do*>.

Additionally, we administered two scales from the German version of the Five Facet Mindfulness Questionnaire, FFMQ (Baer et al., 2008; Michalak et al., 2016). We deemed mindfulness, as the capability for being consciously aware or present in a given moment and for taking account of the currently prevailing situation, as a potentially relevant mediator for the relationship between interoceptive sensibility and movement manipulations given its strong link to interoceptive sensibility (Gibson, 2019; Kabat-Zinn, 1990). We selected the subscales *observation* and a*cting with awareness* as they are interesting with regards to interpersonal differences of interoception in a paradigm assessing the susceptibility towards experimentally manipulated head positions and their impact on lexical choice. The *observation* scale measures the tendency to notice and attend to body sensations among other internal or external experiences, containing eight items, e.g., <*When I'm walking I deliberately notice the sensations of my body moving*>. The *acting with awareness* scale contains eight items assessing the attention to momentary activities as opposed to being automatically directed elsewhere, e.g., <*I am easily distracted*>. Participants indicated how often each of the questionnaire statements applies to them on a 5-point Likert scale ranging from *never or very rarely true* (1) to *very often or always true* (5).

Procedure. The procedure was identical to the procedure in Experiment 1 except that participants were given the interoception and mindfulness questionnaires before completing the sentence fragments. Questionnaires were presented in counterbalanced order, the statements within the questionnaires were presented in a randomized order. The tone-movement assignment of Experiment 1 was reversed here such that participants moved their head upwards after a low tone and downward after a high tone.

After the word production phase, participants rated the words which they had produced according to their spatial properties. For this purpose, the experimenter entered the produced words into a spreadsheet during the language production task. Words produced twice were entered only once. Additionally, 20 control words were included in the spreadsheet. Words in this spreadsheet were randomized before being sent to the participants via mail. They were instructed to rate the word referent's spatial locations on a 7-point Likert scale by taking into account the intended meanings during the production phase. For example, we assumed that a word like *cigarette* when being produced after a sentence like *<You are at the bus stop and you see ...>* might be rated as being downward in case the participant had envisioned a cigarette stub lying on the ground. Exploratory analyses of these self-rated words can be found in Supplement 3.

For the main analysis which was based on rating data collected in the same way as in Experiment 1, all produced words were rated according to their spatial properties by a different group of raters (n = 37) who had not taken part in the Experiment 2. Data of one rater were not considered for analysis as data loss exceeded 20 %. Additionally, four raters were excluded due to showing low interrater reliability with the control words calculated using the psych package (Revelle, 2018), $ICC < .75$. Ratings of the remaining raters showed a consistency with previously established spatial location ratings of control words in the range of $ICC = .80$, $95\%\ CI[.62, .80]$ to $ICC = .97$, $95\%\ CI[.94, .99]$ and agreement between raters was excellent according to the criteria of Fleiss (1986), with $ICC = .84$, $95\%\ CI[.77, .91]$. The final data set included 32 raters (23 female, 18 - 76 years, $M_{age} = 23.31$, i$SD_{age} = 4.13$). Each word was rated by at least

15 different persons. Standard deviations for each word over raters varied from $SD = 0$ for the word *Mond*, English: *moon*, to $SD = 2.34$ for *Schnee*, English: *snow*.

Data Analysis

Data preprocessing. In the word production phase, 55 participants completed 42 sentences twice (84 trials) resulting in 4620 data points. Missing data or trials where participants did not follow the instructions were removed, leaving 4510 trials. If less than 80 % of the raters judged the spatial location of a word, they were not considered for subsequent analysis, further reducing the amount of data points for analysis to 4219. Additionally, distributional semantic similarity values between the nouns in the auditorily presented sentences and the produced words were included in the analyses. Cosine values were computed for 4111 trials as not all produced words were included in the corpus and finally, trials with cosine values ≤ 0 were removed from the data set. In total, 4045 data points entered statistical analyses.

Statistical Analyses. Data analysis followed the procedure described for Experiment 1. We first fit a linear mixed effects model containing all the MAIA facets which we had collected in Experiment 2, i. e. in addition to *attention regulation* and *noticing* and their interactions with movement we also added *body listening* and its interaction with movement to the fixed effects in the model. We fit a second linear mixed model where we replaced the interoceptive sensibility covariates from the MAIA by using the facets from the FFMQ questionnaire, i. e. we entered fixed effects for *observation* and *acting with awareness* as well as their interactions with movement as fixed effects. Model random effect structures were reduced following the above procedure before adding the interaction of semantic similarity with sentence location to the fixed effects.

Results

The absolute number of produced words categorised as high ($n_{moveup} = 315$, $n_{movedown} = 290$) or low ($n_{moveup} = 589$, $n_{movedown} = 621$) per movement condition did not differ significantly, $X^2(1) = 1.72$, p = .189. Numerically, small differences in spatial mean ratings of produced words depending on whether participants had performed an upward ($M = 3.77$, $SD = 1.22$) or downward head movement ($M = 3.69$, $SD = 1.20$) were observed. Using a linear mixed effects model (see Table 2 for full model output) we found that the spatial properties of the produced words were significantly influenced by the head movement, i. e. participants produced significantly higher located words when performing an upward head movement compared to a downward movement. None of the interoceptive sensibility measures had a significantly moderating influence on the spatial properties of the produced words although the interaction of movement and *attention regulation* was in a similar effect size range compared to Experiment 1. Similarly, none of the mindfulness measures had a moderating influence on the spatial properties of the produced words (see Supplement 4).

Table 2

Linear mixed model predicting spatial location of produced words by taking the MAIA scores attention regulation, noticing, body listening and their interaction with head movement into account.

Model formula

word location ~ 1 + movement + sentence_location + movement x sentence_location + attention_regulation + attention_regulation x movement + noticing + noticing x movement + body_listening + body_listening x movement + sentence_location x similarity + (sentence_location + attention_regulation x movement || subject) + (attention_regulation x movement + body_listening x movement | stimulus)

Fixed Effects	Estimates [95% CI]	Std. Error	*t*-value	*p*-value
Intercept	3.74 [3.65;3.82]	0.04	83.89	**<0.001**

Movement	0.08 [0.01; 0.16]	0.04	2.19	**0.029**
Sentence Spatial Location	0.28 [0.21; 0.36]	0.04	7.15	**<0.001**
Movement x Sentence Spatial Location	-0.04 [-0.10; 0.02]	0.03	-1.30	0.195
Attention Regulation	0.01 [-0.06; 0.09]	0.04	0.34	0.734
Movement x Attention Regulation	0.14 [-0.03; 0.31]	0.08	1.62	0.105
Noticing	0.00 [-0.06; 0.07]	0.03	0.10	0.920
Movement x Noticing	-0.05 [-0.18; 0.07]	0.06	-0.86	0.390
Body Listening	0.02 [-0.03; 0.08]	0.03	0.86	0.391
Movement x Body Listening	0.04 [-0.07; 0.14]	0.06	0.64	0.524
Sentence Spatial Location x Semantic Similarity	0.35 [0.11; 0.59]	0.12	2.86	**0.004**

Note. Number of participants = 55; number of categories = 42; total n = 4045; CI = confidence interval around the estimate; *p*-values are based on Satterthwaite approximation as implemented in lme4. Significant p-values of p < .05 are shown in bold.

Furthermore and mirroring the effects from Experiment 1, vertical spatial sentence locations influenced the spatial properties of the produced words significantly with increasingly higher vertical positions of the produced words with increasing vertical position of the presented sentences. There also was a significant interaction between semantic similarity and sentence location with more pronounced effects of spatial locations of the presented sentences on spatial properties of the produced nouns for higher semantic similarity between presented sentence and produced noun.

Discussion

The pattern of results closely resembles the pattern of results from Experiment 1. We found a significant effect of head movement on lexical selection, i. e. participants' word choices were influenced by the head movement direction with a linear increase in spatial location of noun referents for head up compared to head down movements. Probably, and in line with the aim of this manipulation, the reversed order of testing in comparison to Experiment 1 with participants first answering the interoception and mindfulness questionnaires before conducting the sentence completion task enhanced the sensitivity towards the bodily manipulations thus strengthening the effect of head movement on lexical selection across participants. This interpretation is further corroborated by the lack of an interaction between the interoceptive sensibility measure *attention regulation* with the head movement condition in Experiment 2 while there was an interaction in Experiment 1. The effect of head movement on lexical selection in Experiment 1 may have been more strongly affected by differences in interoceptive sensibility, whereas in Experiment 2, the questionnaire conducted before sentence completion may have induced a momentarily stronger focus on interoception across participants, levelling out interindividual differences.

Furthermore, we can rule out that the effects found in Experiment 1 can be traced back to a semantically compatible tone-response-assignment. In Experiment 1, upward head movements always followed a high tone and downward head movements always followed a low tone. However, a main effect of head movement was observed in Experiment 2 even though the link between the tone and the response direction was more opaque.

Pooled analysis

As planned and pre-registered, the data from Experiment 1 and 2 where semantic similarity values had been obtained was pooled. Considering all trials from Experiment 1 and 2 regardless of whether semantic similarity values were obtained yielded similar results, see Supplement 5. The pooled data set consisted of 7199 observations from 99 participants. Data analysis followed the procedure described for Experiment 1, i. e. interoceptive sensibility measures as indicated by the *noticing* and *attention regulation* scale from the MAIA were included as these measures had been obtained in both experiments. Additionally, we included fixed effects for experiment as well as interactions between experiment and all the fixed effects from the previous analyses. We started with the same maximal random effect structure as for Experiment 1 and simplified the random effect structure as outlined above. As none of the interactions with experiment showed significance, we checked whether leaving out those interactions would result in a significant loss in goodness of fit. As this was not the case, all interactions were dropped, only the main effect for experiment remained in the model. At the end, we added the interaction between semantic similarity and the spatial preratings of the presented stimulus sentences to the fixed effects, see Table 3 for output of the final model.

Table 3

Linear mixed model predicting spatial location of produced words by taking the MAIA scores attention regulation, noticing and their interaction with head movement into account using pooled data from Experiment 1 and 2.

Model formula

word location ~ 1 + movement + sentence_location + movement x sentence_location + attention_regulation + attention_regulation x movement + noticing + noticing x movement + experiment sentence_location x similarity + (sentence_location + movement x sentence_location || subject) + (attention_regulation x movement || stimulus)

Fixed Effects	Estimates [95% CI]	Std. Error	*t*-value	*p*-value
Intercept	3.65 [3.56; 3.74]	0.05	78.31	**<0.001**
Movement	0.08 [0.03; 0.13]	0.03	2.94	**0.003**
Sentence Spatial Location	0.29 [0.21; 0.37]	0.04	7.03	**<0.001**
Movement x Sentence Spatial Location	-0.02 [-0.06; 0.04]	0.03	-0.30	0.767
Experiment	0.17 [0.11; 0.23]	0.03	5.64	**<0.001**
Attention Regulation	0.02 [-0.03; 0.07]	0.02	0.82	0.414
Movement x Attention Regulation	0.13 [0.04; 0.22]	0.05	2.76	**0.006**

Noticing	0.00 [-0.05; 0.04]	0.02	-0.18	0.858
Movement x Noticing	-0.07 [-0.16; 0.01]	0.04	-1.71	0.088
Sentence Spatial Location x Semantic Similarity	0.48 [0.30; 0.66]	0.09	5.17	**<0.001**

Note. Number of participants = 99; number of categories = 42; total n = 7199; CI = confidence interval around the estimate; *p*-values are based on Satterthwaite approximation as implemented in lme4. Significant p-values of p < .05 are shown in bold.

We found a main effects for head movement, with words being higher up after upward compared to downward head movements as well as a main effect of spatial location of sentences with the spatial properties of produced words being influenced by the spatial characteristics of the auditorily presented sentences (see Figure 2A). Mirroring the results of Experiment 1, the interaction between head movement and the interoceptive sensibility score *attention regulation* became significant with participants scoring higher on the scale of *attention regulation* being more influenced by the head manipulation. Additionally, there was a marginally significant interaction between head movement and the interoceptive sensibility score *noticing* which taps into participants' awareness of uncomfortable, comfortable and neutral body sensations (see also Supplement 1). People with higher scores of *noticing* showed a tendency to be less influenced by the head movement manipulation (see Figure 2B).

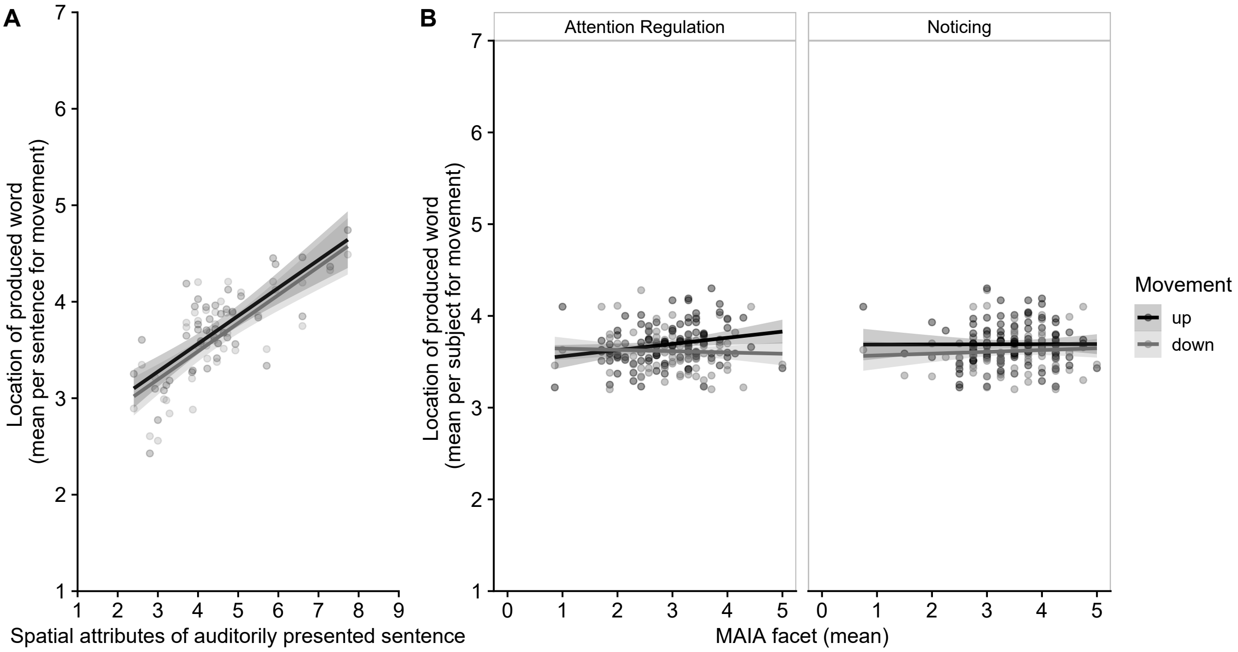

Figure 2. Based on the pooled dataset, Figure 2A shows differential effects of head movement on spatial properties of produced words and Figure 2B depicts the interaction of head movement with the interoceptive sensibility facets *attention regulation* (left) and *noticing* (right). Participant with higher attention regulation abilities showed an effect of head movement on the spatial properties of the produced words.

Furthermore, the main effect of sentence location on the spatial properties of the produced noun was significant as well as the interaction of the spatial properties of the presented sentence with the semantic similarity of produced noun and noun in the stimulus sentence. This indicates that the spatial properties of the sentence material had a higher influence the higher the semantic similarity between a sentence and its chosen ending (see Figure 3). There also was a significant main effect of experiment, mirroring the difference in intercepts between Experiment 1 and 2, see Table 3. In general, pooling stabilized the effects we found in the analyses for Experiment 1 and 2.

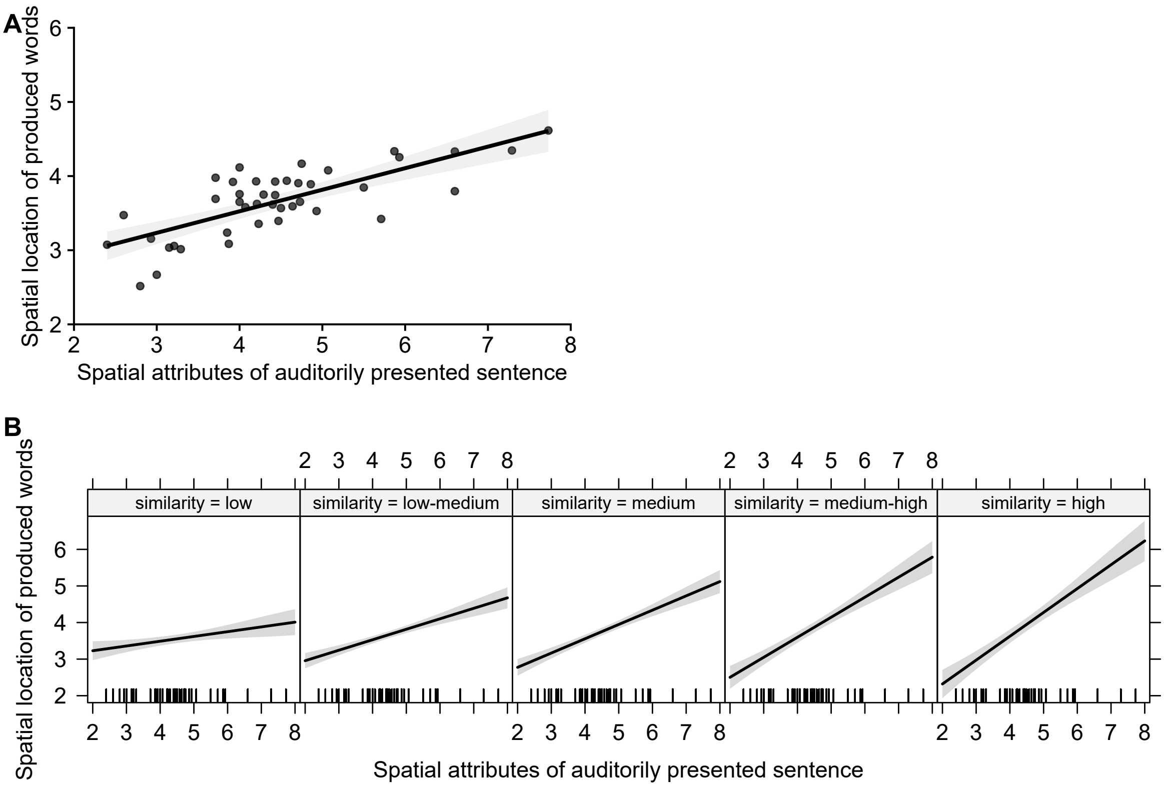

Figure 2. Effect of sentence spatial location on spatial location of produced nouns as a main effect (3A) and in interaction with different degrees of semantic similarity between the sentence noun and the produced noun (3B) for the pooled data set. Higher values for location of the produced noun – as well as for spatial attributes of the written sentences – indicate a higher localization in space. For Figure 3B, small ticks above the x-axis mark the spatial property distribution of the set of sentences. For illustrative purposes, sentence spatial locations are not centered as they were in the analysis. The continuous predictor of similarity was split into five points of equal distance. Low and high similarity refer to the lowest vs. highest cosine values obtained in this study. They are used as descriptive labels, while no pre-defined level of degrees of semantic similarity with regard to cosine values exists.

General Discussion

In two experiments we investigated whether changes in body posture influence lexical choices in a non-constraining sentence completion task. Participants were asked to complete auditorily presented sentences while performing upward or downward head movements. The spatial attributes of the produced nouns were subsequently rated by an independent group of participants. Across both experiments we found that the words which were chosen as sentence endings were influenced by the head movements: the spatial location of a word's referent was higher up in space after an upward head movement and lower in space after a downward head movement. This effect was either observed as a main effect (Experiment 2) or in interaction with individual differences in interoceptive sensibility (Experiment 1). In a combined analysis of the two experiments, the main effect as well as an interaction were found. The higher participants' tendency to control and pay attention to bodily sensations, the more their lexical choices were influenced by the direction of head movements. We also replicated previous findings that the spatial characteristics of produced nouns are influenced by the spatial properties of the presented sentences (Vogt, Kaup, & Abdel Rahman, 2022).

Sensorimotor activations and their impact on lexical selection

In line with accounts of Embodied Cognition our findings indicate that head movements reactivate experiential traces of space which then affect the selection of words from among other lexical candidates. Similar effects of body posture have been obtained by Dijkstra et al. (2007) who found that certain autobiographical memories are easier to retrieve when people are in a congruent body posture, e.g., thinking about the dentist when lying on the back. Furthermore, Lachmair and colleagues (2016) found that word recall for up- and down-related words was better when participants where in a congruent body position. We extend these findings by showing that body posture also has an effect on language production and the words we select during speech planning.

Previous research from our own lab showed that minimal reactivations of experiential traces of space by visually presenting the sentence fragments in an upward or downward moving direction did not lead to a comparable effect (Vogt, Kaup, & Abdel Rahman, 2022). In contrast, the more direct and possibly stronger manipulation of bodily activation in this study was sufficient to reactivate experiential traces of space. These observations are in line with previous work showing that the experimental manipulations which are meant to reactivate experiential traces need to be relevant and meaningful enough in order to influence semantic processing (Berndt et al., 2018; Ostarek & Vigliocco, 2017).

Sensitivity to internal body sensations affects embodied lexical selection

Sensorimotor experiential effects on lexical selection may be enhanced when people focus on bodily sensations or when they have a general tendency to be sensitive. Therefore, we additionally investigated whether interindividual differences in different facets of interoceptive sensibility, the sensitivity to processes originating inside or concerning our bodies (Häfner, 2013), would enhance the effects of sensorimotor influences on lexical selection. We found that the effect of head movement on spatial properties of produced nouns was more prominent the higher participants scored in *attention regulation*, a measure of the ability to sustain and control attention to body sensations (Mehling et al., 2012). The interaction of individual characteristics of interoceptive processes with sensorimotor experiences may be one of the forces determining how well we can verbally express our sensations to others, as a key human capability. This is one of the first pieces of evidence that individual cognitive traits are among the factors influencing lexical choices during speaking.

On a general note, this finding highlights the role of interindividual differences in cognitive processing which have not gained much attention in research on lexical access during speaking so far. Possibly, interindividual differences may also be contributing to the replication crisis in the field of psychology and may be among the reasons for non-replicable effects or opposing outcomes in the domain of embodied language processing more specifically (Morey et al., 2022; Ostarek & Huettig, 2019). Therefore, we hope to inspire new research addressing the role of interindividual differences in language processing in the future.

The interaction between head movement and the interoceptive sensibility measure *attention regulation* was significant in Experiment 1 and in the pooled analysis, but did not reach significance in Experiment 2. Our account for this pattern of results is that interindividual differences in this facet were alleviated in Experiment 2 as participants filled out the questionnaires on interoceptive sensibility before the sentence completion in Experiment 2. Thereby, and as intended by this manipulation, all participants might have been paying more attention to their body sensations in comparison to Experiment 1, leveling the impact of interindividual differences in this domain. This is also reflected in differences in between-subject variability in *attention regulation*, which were generally smaller in Experiment 2 ($SD_{AR\_Exp2} = .66$) than in Experiment 1 ($SD_{AR\_Exp2} = .83$), and which may have contributed to the observed pattern.

We also observed a trend for an interaction between the interoceptive sensibility score *noticing* and head movement in the pooled analysis as well as the first part of Experiment 1. While both *attention regulation* as well as *noticing* capture the sensitivity of participants towards their own bodies, *noticing* focusses on the awareness of uncomfortable, comfortable and neutral body states (Mehling et al., 2012). However, future work is needed to reveal whether this observation is replicable before drawing conclusions.

As, to the best of our knowledge, no previous research on the role of interoception in language production is available, we also explored other scales which tap into traits connected to interoception. However, neither of these modulated the effects of head movements on lexical selection.

Thus, the scales *attention regulation* and possibly *noticing* from the MAIA seem the most promising candidates for assessing interindividual differences in interoception and their relationship with language production. Potentially, they can be complemented with measures of interoceptive accuracy which do not only rely on self-report (e.g., the ability to correctly infer ones heart beat) in the future.

Experiential traces of space embedded in linguistic context influence lexical choices

As a secondary goal, we replicated the results from a previous study (Vogt, Kaup, & Abdel Rahman, 2022), showing that the broader spatial context of the situations described in the sentence fragments influenced the spatial properties of the words which had been chosen as suitable sentence endings. The more a sentence referred to a situation in the upper or lower domain of the world, the higher up or lower down the referents of the produced nouns were located. This interacted with a distributional measure of semantic similarity (Günther et al., 2015) such that the spatial effects were stronger for semantically close noun-sentence combinations. However, even in cases of minimal semantic similarity the effect of sentence spatial location on the spatial properties of the produced nouns was present, indicating that the effects of sentence spatial location cannot be explained entirely by the semantic relationship between the presented sentences and the chosen sentence endings. This is in line with theoretical accounts and recent findings (Banks et al., 2021; Louwerse, 2018; Vigliocco et al., 2009). These robust findings can be explained in the context of hybrid theories of semantic memory which integrate experientially grounded meaning aspects and usage-based distributional aspects of language as separate but interacting types of meaning (Bi, 2021; Carota et al., 2021; Davis & Yee, 2021).

We take this as evidence that lexical choices are influenced by the experientially grounded sensory meaning of space not only by manipulation of bodily movements but also as conveyed by the verbal context. Note, that we did not find an interaction between body movements and sentence spatial locations, suggesting that different aspects of meaning related to sensory experiences grounded in our experiences are influenced by these manipulations which is in line with hybrid models of semantics.

A causal role of embodied meaning for language production

In comprehension research it has been shown that the presentation of words with motor content is linked to motor priming or motor cortex activations (Yee et al., 2013)). It has been debated whether such reactivations of embodied meaning are relevant for comprehension or may be viewed as mere by-products of our learning history without a direct relevance for the reconstruction of meaning (Kaup et al., 2016; Pulvermüller, Hauk, Nikulin, & Ilmoniemi, 2005; Strozyk et al., 2019; Willems, Labruna, D'Esposito, Ivry, & Casasanto, 2011). Probing causality has been highlighted as one of the challenges for investigations of experientially grounded meaning in the future (Ostarek & Bottini, 2020; Ostarek & Huettig, 2019).

In contrast, language production is well-suited to address questions on the causal impact of embodied meaning on language processing. It can be tested whether grounded meaning has an effect on the co-activation of alternatives to an intended message (see e.g., Hirschfeld & Zwitserlood, 2012) or, as done here, by testing whether grounded meaning activations may even influence which words we select for speaking. Based on the results from Experiment 1 and 2, we conclude that embodiment in the form of activations of sensorimotor experiences can directly affect which words we chose in an open language production task, an outcome of high theoretical relevance.

Limitations

Alternatively, could the effects be due to participants internally verbalizing the direction in which they moved their head (e.g., UP vs. DOWN), which may lead to the activation of concepts which share these semantic features? This seems unlikely since it has been shown that spatial compatibility effects in language comprehension are still evident even if no internal response labelling occurred (Dudschig & Kaup, 2017). To directly rule out this alternative explanation, future studies should also assess participant's tendency to verbalise internally. Another potential confound relates to the execution of head movements and hypotheses of participants about the aims of the study, which could have caused them to intentionally produce words higher or lower in space. However, this is also not likely because few have correctly inferred the goal of the study (n = 9 across both experiments), and those participants were excluded. Furthermore, such a strategy could not explain the interaction with interoceptive sensibility that we observed. We are therefore confident that our findings reflect effects of embodiment on lexical selection.

Conclusion

The present findings demonstrate that sensorimotor experiential activations may directly be reflected in the words we chose when we speak and that increasing levels of sensitivity to bodily sensations enhance this effect, highlighting the role of individual differences for lexical-semantic processes during speaking: how we verbally express our sensations and feelings depends on sensorimotor experiences and on our sensibility to access these experiences.

While most paradigms investigating semantic processing in language production use picture naming tasks in which lexical selection is constrained, we employed a sentence completion task in which lexical selection is relatively free. With this task we demonstrate that (embodied) meaning shapes not only lexical-semantic co-activation and naming latencies, but also affects the words we chose to verbally express ourselves, extending our knowledge on embodied cognition and language production. Despite being a small effect, its communicative consequences might be large because entirely different lexical items might be chosen depending on the bodily state a person is in and on their access to this state.

Acknowledgements

Anne Vogt was supported by a doctoral fellowship of the German National Academic Foundation (Studienstiftung des deutschen Volkes) and the Berlin School of Mind and Brain. We thank Nora Holtz for recording the stimulus material and Ilan Nestic for collecting the data for Experiment 2. All authors developed the study design, discussed the results and wrote the manuscript. Anne Vogt and Isabel Ganter programmed the experiment and performed the analyses and Isabel Ganter collected the data for Experiment 1.

3.3 Manuscript 3: Internet based language production research with overt articulation: Proof of concept, challenges, and practical advice

This manuscript was published as:

Vogt, A.*, Hauber, R.*, Kuhlen, A., & Abdel Rahman, R. (2022). Internet-based language production research with overt articulation: Proof of concept, challenges, and practical advice. *Behavior Research Methods, 54*(4), 1954-1975. https://doi.org/10.3758/s13428-021-01686-3
* equal contribution

This is the final author copy with small formatting adjustments for inclusion in this dissertation. Data and analysis scripts as well as demo files for online audio recording are available in a data repository: https://osf.io/uh5vr/

References are listed at the end of the dissertation.

Abstract

Language production experiments with overt articulation have thus far only scarcely been conducted online, mostly due to technical difficulties related to measuring voice onset latencies. Especially the poor audiovisual synchrony in web experiments (Bridges et al., 2020) is a challenge to time-locking stimuli and participants' spoken responses. We tested the viability of conducting language production experiments with overt articulation in online settings using the picture–word interference paradigm – a classic task in language production research. In three pre-registered experiments (n = 48 each), participants named object pictures while ignoring visually superimposed distractor words. We implemented a custom voice recording option in two different web experiment builders and recorded naming responses in audio files. From these stimulus-locked audio files, we extracted voice onset latencies offline. In a control task, participants classified the last letter of a picture name as a vowel or consonant via button-press, a task that shows comparable semantic interference effects. We expected slower responses when picture and distractor word were semantically related compared to unrelated, independently of task. This semantic interference effect is robust, but relatively small. It should therefore crucially depend on precise timing. We replicated this effect in an online setting, both for button-press and overt naming responses, providing a proof of concept that naming latency – a key dependent variable in language production research – can be reliably measured in online experiments. We discuss challenges for online language production research and suggestions of how to overcome them. The scripts for the online implementation are made available.

Keywords: Language production, voice onset latency, online experiments, overt articulation, picture word interference

Reasons for conducting online language production experiments

Many psychological experiments based on behavioral measures can be run online. This brings great advantages in comparison to lab-based testing. For example, online experiments facilitate testing larger samples, promote science to a larger community, and potentially consume less resources during data collection (e.g., Grootswagers, 2020). This greater efficiency in data collection has led to the replication and extension of many behavioral paradigms in online settings. Even experiments which require precise measures of reaction times can reliably be conducted on the web (e.g., Anwyl-Irvine, Dalmaijer, Hodges, & Evershed, 2021; Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020; de Leeuw, 2015; Gallant & Libben, 2019; Hilbig, 2016; Pinet et al., 2017). Furthermore, there is an increasing awareness for the need to test more diverse populations in order to raise the external validity of experimental findings (Speed, Wnuk, & Majid, 2018). Web-based testing is one of the options to ensure that our understanding of the human mind extends to the population at large. Most recently the popularity of web-based testing has been gaining additional momentum as the pandemic Covid-19 forces researchers to think of alternatives to lab-based testing (Sauter, Draschkow, & Mack, 2020).

Within psycholinguistics, language production research is so far underrepresented when it comes to online-based testing. To the best of our knowledge, typical language production tasks such as picture naming, during which participants' overt articulatory responses are acquired in order to determine voice onset latencies, have so far not been investigated in online settings. In this study, we provide a proof of principle for deriving voice onset latencies from recordings of overt articulatory responses time-locked to pictorial stimuli. To this end, we implemented the picture word interference (PWI) task, a classic paradigm to investigate lexical access during language production (for a recent review see Bürki et al., 2020), in an online version. We demonstrate the viability of this approach by comparing voice onsets computed from short audio recordings of overt naming responses with a manual classification task providing response times depending on manual keyboard button-press responses. In a direct comparison, we show similar interference effects typically observed in lab-based picture word interference studies for both overt picture naming responses and for manual button-press classifications of the picture names. Furthermore, we provide practical advice on moving language production research online.

Current challenges when running language production studies online

There are several challenges to conducting online language production research relying on overt naming responses. First, in the lab, the technical equipment (e.g., microphones, sound shielded booths) ensures a high quality of the acquired speech data and technical requirements are kept constant within and across participants. In an online study, participants need a microphone and need to explicitly grant access in order to record speech as dependent variable. This approval process can disrupt the experimental procedures at different time points – depending on the individual browser's security settings. Furthermore, recording quality may differ widely between participants due to technical reasons or background noise. Second, in the lab, the experimenter would typically monitor if the participant is complying to the instructions, e.g., naming the pictures presented on the screen in a correct manner. However, there is no easy way to monitor the performance of participants' online verbal responses. In an online language production experiment, recorded verbal responses can only be checked after completion of the experiment. Third, lab-based experiments have the option of using voice keys, special hardware devices for automatically detecting the onset of vocal responses. Alternatively, or in addition, in the lab vocal responses can be recorded as audio files, which are scanned for the start of speech in a subsequent step after the experiment. Voice onset latencies are a key dependent variable in many language production experiments and result as the latency between the onset of a picture presentation and the onset of the naming response. However, none of the available tools for running online experiments offers the possibility to determine voice onset latencies instantly and directly log them rendering online language production research potentially more laborious after data acquisition.

Lastly, and perhaps the most serious challenge, is to precisely timelock vocal response to certain events, e.g., the visual presentation of a picture on a screen, with little or no variation within and between participants. In the lab, experimenters use specific hard- and software to control for audiovisual synchrony ensuring that the timing between stimulus presentation and voice recording is reliable. Unfortunately, to date none of the packages or programs for running online experiments offer an option for recording overt articulatory responses precisely time-locked to a (visual) stimulus. Only very recently there has been some development in this area resulting in beta versions of experimental software with audio recording possibilities and there seems to be a lively development process surrounding these versions (e.g., see the Gorilla Audio Recording Zone, or the *jspsych-image-audio-response-plugin* by Gilbert, 2020). However, none of these versions can to date ensure the high audiovisual synchrony which would be needed in order to test the typically investigated effects in language production research, which sometimes rely on mean voice onset differences in the range of a few miliseconds. To date, a published validation of their timing properties is still missing.

A recent meta-analysis compares a range of experiment builders concerning the reliability of synchronous presentation of visual and auditory stimuli, both lab-based and online, and testing different operation systems and browsers (Bridges et al., 2020; Reimers & Stewart, 2016). These studies demonstrate that the lag between visual and auditory onset varies considerably. As audio output timing, i. e. the start of audio recordings, relies on specific hardware and software properties as well, it can be assumed that it is likewise difficult to control for a precisely stimulus-locked onset of an audio recording in online settings where a large number of participants with different computer and browser configurations takes part. Bridges et al. (2020) conclude that for online experiments, none of the tested packages can guarantee reliable audiovisual synchrony. They argue that JavaScript technology, which is the basis for most web-based experiments, would need to be improved in order to obtain precisely timed audio measures in the millisecond range. However, precisely timelocking overt articulation to pictorial stimuli might be an essential prerequisite to ensure replicability of language production paradigms in an online setting (Plant, 2016).

For the current study, as a proof of concept, we adopted a pragmatic approach to these challenges. As Bridges and colleagues (2020) point out, solving the problem of poor audio-visual synchrony in online software technology is not a task for the scientific user but rather for the community of software developers working with JavaScript. In the meantime, however, we aim for a “good-enough” approach, i. e. answering the question if the current methods are reliable *enough* to detect mean differences in classic paradigms and to replicate classic effects. Even for measurements made with instruments with poor resolution, mean differences can be successfully detected when aggregating over a large enough sample size (Ulrich & Giray, 1989). Therefore, given a sufficiently large number of trials and participants, one might be able to detect differences in naming latencies recorded in online settings in spite of the multiple challenges and limitations of web technology (Brand & Bradley, 2012). The same approach applies, to some extent, more broadly to all experimental research conducted online: while the data collected in these settings will invariably show some increased noise due to lack of experimental control and limitations introduced by variable personal hardware (e.g., laptop keyboards), effects should still be, and indeed are, detectable with sufficient power (Mathot & March, 2022; Pinet et al., 2017).

Testing an online implementation of the picture word interference paradigm with verbal and manual responses

Given the many advantages of online experiments, the aim of the present study is to test whether a robust and well-replicated, but relatively small effect in language production can be replicated in an online implementation of the task including overt naming responses. In the picture word interference paradigm, participants name pictures while ignoring simultaneously presented distractor words. Naming latencies in this paradigm depend on the semantic relation between the distractor and the target picture, with increased naming times for semantically related versus unrelated distractor words (e.g., Lupker, 1979; Schriefers et al., 1990). This well-replicated semantic interference effect has been interpreted as a marker for the cognitive processes underlying lexical access (Bürki et al., 2020). Analysing results from 162 studies with a Bayesian meta-analysis, Bürki and colleagues demonstrated that the semantic interference effect amounts to 21 ms with a 95 % credible interval ranging from 18 to 24 ms. Therefore, replicating this small but robust effect in an online setting would demonstrate the viability of running time sensitive language production experiments online.

Crucially, semantic interference has not only been demonstrated for overt naming responses. A comparable effect has also been observed for a manual button-press classification task in which participants are asked to identify the last letter of the picture name as a vowel or a consonant by pressing one of two response buttons (Abdel Rahman & Aristei, 2010; Hutson et al., 2013; Tufft & Richardson, 2020). The similarity of semantic interference in vocal and manual naming responses allows us to directly compare these effects in an online implementation of the paradigm. While manual responses can more readily be implemented and recorded in online settings, we can compare the effects in this response modality directly to the recording of overt naming responses.

To this end, we adopted the design of the study by Abdel Rahman and Aristei (2010) in which participants received both versions of the task with the experimental manipulation of semantic relatedness as a within-subject factor. In this way, manual button response times can serve as a benchmark for a potential effect in the vocal onsets in audio responses. For naming latencies, we recorded audio files for each naming trial, time locked to the onset of picture presentation. Naming latencies were then computed offline. To enable online voice recording, we customized freely available tools for audio recordings on the web and included them in two different experiment builder programs. It was our primary goal to find a working solution for running language production experiemnts online. Therefore, we conducted the same experiment in two different implementations to increase chances for finding a working solution.

The first version was programmed and hosted in SoSciSurvey (Leiner, 2019), a platform used for conducting social and behavioral research in Germany in combination with an audio recording function based on RecordRTC (Khan, 2020). The second implementation was programmed using jsPsych (de Leeuw, 2015) with a custom audio recording plugin relying on recorder.js (Diamond, 2016) and hosted on JATOS (Lange et al., 2015). The methods and predictions of this work have been preregistered on AsPredicted.com (AsPredicted#: 43871, available for viewing under https://aspredicted.org/blind.php?x=6ma52w). Data and analysis scripts are available at OSF (https://osf.io/uh5vr/?view_only=229679aa33604aa2a5cb400eab62099). The comparison between the two implementations is an exploratory analysis which was not preregistered.

Experiment 1

Methods

One version of the experiment was implemented in SoSciSurvey (Leiner, 2019), the other version was implemented in jsPsych (de Leeuw, 2015). The two versions of the experiment (labeled *SoSciSurvey* and *jsPsych1*, respectively) were nearly identical regarding design and procedure. If not specified otherwise the information applies to both versions.

Participants. In the *SoSciSurvey* version, a total of 116 native German speakers between 18 and 35 years were recruited over the commercial platform Prolific (www.prolific.co.uk) and completed the experiment. They were included in the final sample when meeting all inclusion criteria until the final sample consisted of 48 participants as determined in the preregistration (21 females, 18 - 33 years, M_{age} = 25.71, SD_{age} = 4.28). See also the section on Data Exclusion in Data Analysis for details on the criteria for inclusion in the final sample. The sample size was determined via an a-priori power analysis using the simr package (Green & MacLeod, 2016). Simr uses simulation to estimate power, by simulating data for which the user can define the parameter estimates. We estimated the power for the overt naming task but needed to rely on estimates from a study employing the PWI paradigm (Lorenz, Regel, Zwitserlood, & Abdel Rahman, 2018) that used LMMs to analyse their data, which Abdel Rahman and Aristei (2010) did not. The resulting suggested sample size for a power estimate of 80 % was 36, but we anticipated a need for more power in online studies and had decided a priori to increase the estimated sample size by one third, thus amounting to the sample size of n = 48.

In the *jsPsych1* version, a total of 108 native German speakers between 18 and 35 years were recruited over the commercial platform Prolific (www.prolific.co.uk) and completed the experiment. They were included in the final sample when meeting all inclusion criteria until the final sample consisted of 48 participants as determined in the preregistration (24 females, 18 - 33 years, $M_{age} = 26.06$, $SD_{age} = 3.99$).[8]

Participants provided informed consent to their participation in the study. The study was conducted on the basis of the principles expressed in the Declaration of Helsinki and was approved by the local Ethics Committee. Participants received monetary compensation distributed via the platform Prolific.

Materials. The stimulus set consisted of 40 black and white line drawings of common objects, all of which have frequently been used in lab-based picture naming studies in our group. Half of the German words for these objects ended in a vowel, the other half in a consonant. For the related condition each drawing was assigned a semantically related distractor word that was not part of the response set. For the unrelated condition the same distractor words were reassigned to different drawings to which they were not semantically related. In both conditions half of the assigned distractors matched the name of the drawing with regards to the type of the last letter (vowel vs. consonant) and the other half did not. Thus, response compatibility of the distractor and the target regarding the classification of the last letter was balanced across the stimulus set. In the case of a compatible match, it was ensured that the last letters were

[8] When asked whether German was their native language, one participant selected “No” while she had given the information of being a German native speaker on the recruiting platform Prolific. As careful screening of her audio files did not give any hint that German might not be her native language, we decided to keep this participant.

never identical, but only matched concerning the type of letter, i. e. vowel or consonant. The line drawings were presented together with the visual distractor words that were superimposed without obscuring the visibility of the object. See online Supplementary Material A for a table of the stimulus set (https://osf.io/uh5vr/?view_only=229679aa33604aa2a5cb400eab62099).

Design. The experiment consisted of a 2 x 2 design with the within-subject factors *task* (button press vs. overt naming) and *relatedness* (related vs. unrelated distractors). The dependent measure was response latency (of the button press and the overt naming, respectively). The order of the tasks (button press—overt naming vs. overt naming—button press) and the assignment of buttons to responses (*p* for vowel, *q* for consonant vs. *p* for consonant, *q* for vowel) was counterbalanced across participants.

Procedure. At the start of the experiment participants were given general instructions and then a preview of all 40 drawings with the corresponding names (but *without* distractors) to familiarize participants with the stimulus set. Instructions for the first task were then presented including a catch trial to ensure participants read the instructions carefully, followed by four practice trials. Each main task consisted of 80 trials showing all 80 stimuli in random order. Each trial started with a 500 ms fixation cross at the center of the screen. The stimulus, a line drawing with a superimposed distractor word, then appeared at the center of the screen (200 x 200 pixels) for a total of 2000 ms, followed by a blank screen for another 1000 ms before the next fixation cross appeared. In the button press task response labels (e.g., “Q consonant”, “P vowel”) were shown below the stimulus to the left and right respectively. Once a button was pressed, the corresponding response label was highlighted but the stimulus remained on screen for the full duration of the trial. In the overt naming task, audio files were recorded for each trial starting at stimulus onset, producing 80 recordings with a predefined duration of 3000 ms for each participant. There was a short break after the first task before the instructions of the second part were presented. When both tasks were completed, participants received debriefing information and were then linked back to the website of Prolific in order to validate their participation.

Technical Implementation of audio recording. The technical implementations for both experimental platforms relied on JavaScript. JavaScript is a programming language that forms, together with HTML and CSS, the core technology of the internet. Importantly, all modern browsers rely on JavaScript and therefore no prior installation of the language itself is necessary neither on the programmers' nor the users' side. The implementations in our study build on APIs (application programming interfaces) which can be thought of as ready-made tool sets allowing for certain functionalities to be used. The access to participants' microphones and the streaming of their voice input is realized via such APIs in both implementations.

The experimental platform SoSciSurvey (Leiner, 2019) is based mainly on the programming language PHP, but JavaScript code can be implemented in the functionalities provided by SoSciSurvey, as we did in our study. For the implementation of the audio recording we included a JavaScript based function within each audio trial. This function captures the participants' audio input, presents the visual stimulus and starts an audio recording at the same time. The audio input is then saved in the browser's native file format (e.g., .ogg or .webm) and transferred to the SoSciSurvey server. The JavaScript plugin *RecordRTC.js*, which we used for this purpose, is provided by Khan (2020) who provides and actively maintains a wide range of readymade JavaScript applications under the open WebRTC (web real time communication) standard.

In our jsPsych-version the functionality of the experiment timeline relies on the experiment library jsPsych (de Leeuw, 2015) while the data is saved via a server specified by the experimenter, in our case a server based at our institute, set up with JATOS (Lange et al., 2015). For the technical implementation of audio recording within jsPsych, access to participants' microphones is granted only once at the beginning of the overt naming part and remains permanently active during the overt naming task. The recordings are started within each trial upon stimulus presentation. Then files are immediately transferred in wav-format and transferred to the server. Unlike the SoSciSurvey implementation the recording in jsPsych relies on the JavaScript plugin *recorder.js* provided by Diamond (2016) which we used to customize a jsPsych-plugin to enable audio recording. Note that even though the functionality of *recorder.js* builds the base of *RecordRTC.js* (as implemented in the SoSciSurvey implementation described above), and also of other audio recording plugins, it is not actively maintained and therefore might not be working in the future, e.g., if browser standards change.

The main difference with regard to the implementation of the audio recording is that our custom jsPsych-plugin uses a Web Worker API during the recording and saving of audio files. Web Workers allow to run tasks in the background without interfering with the user's interface. This should ensure, for example, that the next trial can start as predefined even if the audio file from the previous trial has not yet been transferred to the server. For anyone interested in more details of the technical background we recommend consulting the MDN Web Docs site as a starting point as this site provides information about Open Web technologies including JavaScript, HTML, CSS, and APIs (https://developer.mozilla.org/de/).

Data Analysis

Data preprocessing. For the *SoSciSurvey* data, the recordings were first converted to wav format from the browser's default compressed recording format. For the *jsPsych* data, the recordings were already in wav format. To extract the naming latency from the audio recordings in the overt naming task, all audio files were then processed with the tool *Chronset* (Roux et al., 2017), which is an automated tool for the detection of speech onsets from audio files. Afterwards, the audio files were manually checked using the software *Praat* (Boersma & Weenink, 2020) to ensure that participants were producing the correct target word and to manually correct the determined speech onset where necessary.

Data exclusion.

Replacement of participants due to prescreening of data. Participants were excluded and replaced in the dataset if more than 20 % of the trials were incomplete or marked as deficient. Trials were marked as deficient if (1) participants produced an error (wrong picture name or wrong letter classification), (2) the audio files of the naming response did not contain any sound, or (3) if there were other technical difficulties concerning the audio files. These difficulties included excessive background noise, an extremely low audio signal, or irregular lengths of the audio file within a participant. In the *SoSciSurvey* version of the experiment, irregular file lengths were so pervasive that we did not consider it practical to replace participants on these grounds in this version. See the Discussion for details on this issue. See also Figure 1 for an overview of the exclusion of participants due to the prescreening criteria.

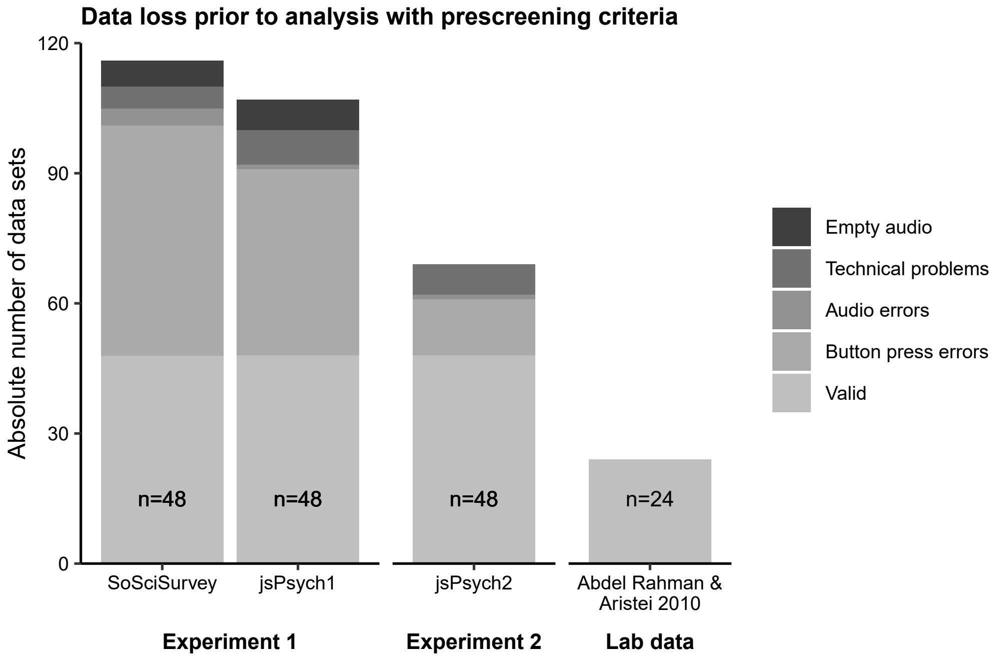

Figure 1. The figure presents the number of individual data sets which had to be collected in order to obtain the pre-defined sample size and the number of data sets that was excluded based on our preregistered inclusion criteria in Experiment 1 and 2. For comparison, the figure also depicts the lab-based experiment from Abdel Rahman & Aristei (2010). In that study, no data sets had to be removed.

Data exclusion of single trials. In the data of the final 48 participants in both versions of the experiment, single trials were excluded if no response was given, participants made an erroneous response, or if participants responded prematurely (i. e. reaction times under 200 ms). See Table 1 for an overview of the data exclusion of single trials.

Table 1

Data loss caused by preprocessing the final samples of n = 48 in % of total data in Experiment 1 (SoSciSurvey and jsPsych1) and Experiment 2 (jsPsych2). Trials were excluded from analysis if participants did not press a button in the binary button press classification task (button press task – no reaction), classified the last letter incorrectly (button press task – error), did not produce an object name in the naming task (naming task – no reaction), did not produce the correct target word in the naming task (naming task – error), or if a voice onset of less than 200 ms was registered (naming task – early response).

	Experiment 1		Experiment 2
	SoSciSurvey	jsPsych1	jsPsych2
Exclusion due to			
Button Press Task – no reaction	1.89	1.48	2.12
Button Press Task – error	3.79	3.92	2.90
Naming Task – no reaction	0.44	0.78	0.20
Naming Task – error	0.69	2.00	1.39
Naming Task – early response	0.12	0.72	0.17
Data Loss	6.93	8.91	6.78

Data transformation and selection of Linear Mixed Effect Models. To approximate a normal distribution of the residuals of the dependent variable, the Box-Cox-power transformation procedure was applied to the response latency data (Box & Cox, 1964). The specific transformation that was performed is noted in the respective section of the Results.

For analysis of the response latency data, we used the packages lme4 (Bates, Mächler, et al., 2015) and lmerTest (Kuznetsova et al., 2017) in the statistical software R (R Core Team, 2019; Version: 3.6.1) to fit Linear Mixed Effect Models (LMMs) of the (transformed) response latency with the fixed effect predictors *task (button press vs. overt naming task)*, *relatedness (related vs. unrelated)*, as well as their interaction. Both predictors were coded as sum contrasts (*button press – overt naming* and *related – unrelated*, respectively). In order to examine more closely the effect of relatedness in the two tasks, nested LMMs were also fitted, in which the fixed effect of *relatedness* was estimated separately for the two levels of the factor *task*. The additional factors *repetition* (whether a picture was seen for the first or second time within a task) and *task-order* (button press—naming vs. naming—button press) were included as separate fixed effects (both contrast-coded). If their inclusion led to an increase in model fit as indicated by a likelihood ratio test, they remained in the final model.

In specifying the structure of the models' random effects, we followed the procedure outlined by Bates, Kliegl, et al. (2015). Initially, a full model with the complete variance-covariance matrix of the random effects allowed for by the design (i. e. random effects by subject and by picture) was fitted. This model was then simplified by first forcing the correlation parameters between the random effects to zero, then identifying overfitting of the parameters in the random effects using principal component analysis and dropping those random effects that contributed the least to the cumulative proportion of variance as identified by the principal component analysis until dropping a random effect led to a reduction in the goodness of fit. Correlation parameters between random effects were then reintroduced and kept in the final model if their re-inclusion led to an increase in the model fit and did not lead to non-convergence of the model. Models reported in the Results section are always final, reduced models.

Results

In the *SoSciSurvey* data, the mean response latency in the button press task was 1161 ms (*SE* = 7 ms) in the related condition and 1143 ms (*SE* = 7 ms) in the unrelated condition. In the overt naming task, the mean response latency was 875 ms (*SE* = 7 ms) in the related condition and 859 ms (*SE* = 6 ms) in the unrelated condition. In the *jsPsych* data, the mean response latency in the button press task, was 1162 ms (*SE* = 7 ms) in the related condition and 1147 ms (*SE* = 6 ms) in the unrelated condition. In the overt naming task, the mean response latency was 1003 ms (*SE* = 6 ms) in the related condition and 989 ms (*SE* = 5 ms) in the unrelated condition. See Figure 2 for line plots of mean response latency by *task* and *relatedness* for both versions of the experiment.

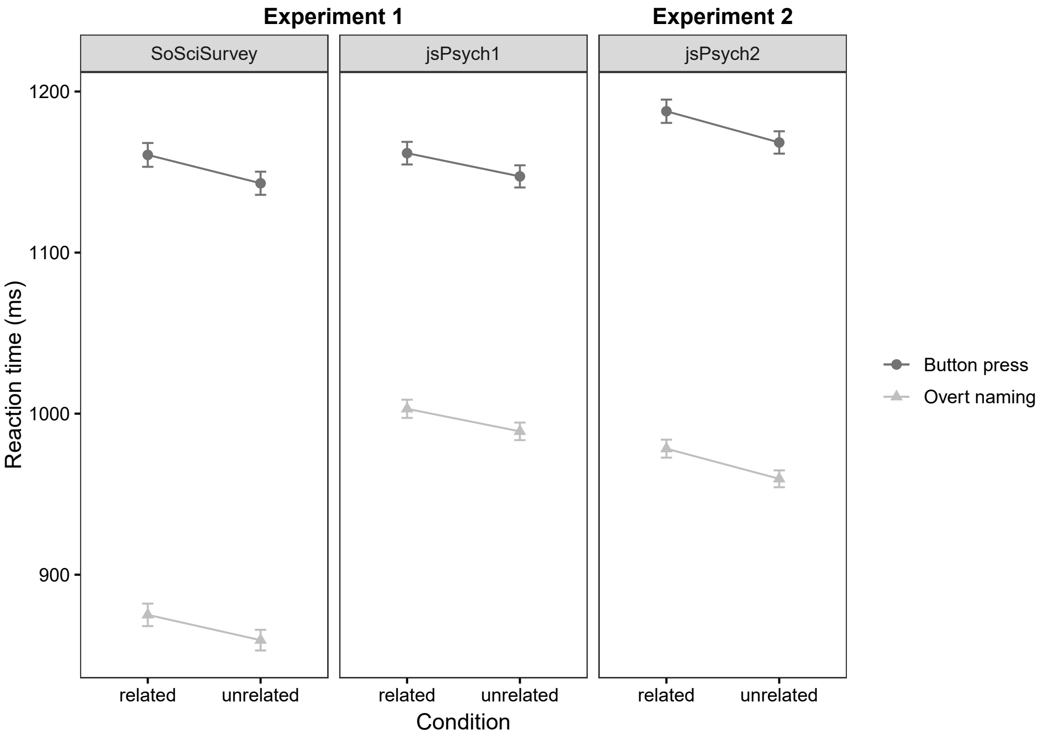

Figure 2. Mean reaction times in ms with standard error of means for naming and button press tasks in both implementations of the online PWI in Experiment 1 (*SoSciSurvey* and *jsPsych1*) and Experiment 2 (*jsPsych2*). Targets presented with a semantically related distractor were classified and named slower than targets with unrelated distractors.

See Figure 3 for line plots of mean response latency by task, relatedness and repetition for all online experiments and Figure 4 for raincloud plots of the single trial data and their distribution by *task* and *relatedness* for all experiments as well as for the previous study by Abdel Rahman and Aristei (2010). Note that overall response latencies are slower in the online experiments compared to the lab. We discuss possible reasons for this in the General Discussion.

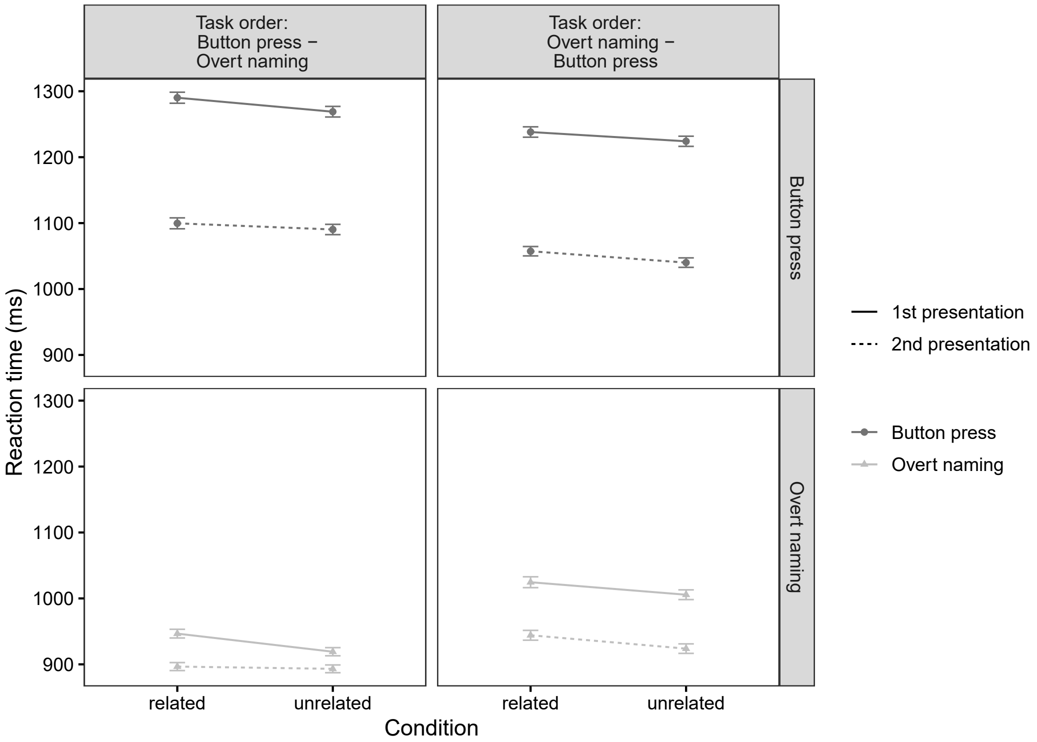

Figure 3. Mean reaction times in ms with pooled data from all online experiments plotted separately for task and task sequence. The figure can be read columnwise for comparing the effect of picture repetition within one task sequence. The left column represents the task sequence 1st button press trials – 2nd overt naming trials and the right column depicts the task sequence 1st overt naming trials – 2nd button press trials. Furthermore, the figure can be read rowwise from left to right in the upper row for comparing the effect of picture repetition (1st-4th) within the button press task and rowwise from right to left in the lower row for comparing the effect of picture repetition (1st-4th) within the overt naming task.

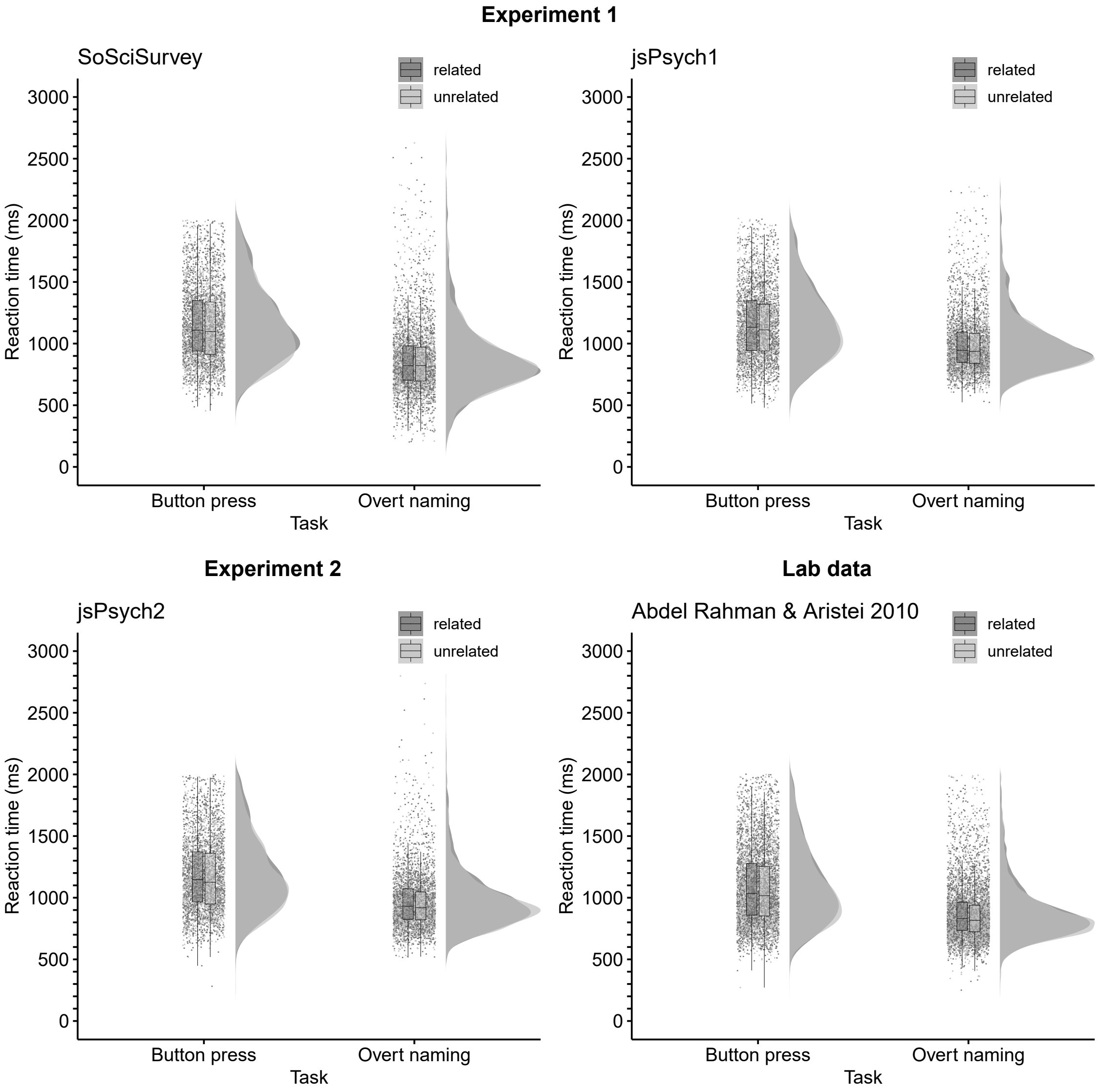

Figure 4. Single trial plots (before model criticism) for the factors *task* and *relatedness* in all three experiments and the lab-based study by Abdel Rahman and Aristei (2010). Box plots represent the median per relatedness condition with lower and upper hinges corresponding to the 25th and 75th percentiles and whiskers extending to the most extreme value within 1.5*IQR from the box hinges.

Preregistered analysis

SoSciSurvey. For the *SoSciSurvey* data, the Box-Cox procedure suggested a log-transformation of the response latency variable. In the final model (containing main effects of *task, relatedness,* and their interaction) both main effects were significant, while the interaction *task*relatedness* was not significant. The positive sign of the estimate for *task* and the coding of the *task* contrast show that response latency was slower in the button press task compared to the naming task. Likewise, for *relatedness* the response latency was slower in the related than in the unrelated condition. In the final nested model (i. e. estimating separate fixed effects of *relatedness*, for the two levels of task), the nested effect of *relatedness* was marginally significant in the button press task and did not reach significance in the overt naming task. See Table 2 for an overview of the reported models from the preregistered analysis including model formula, coefficients and random effect variance parameters.

Table 2

Table of final models from the preregistered analysis of the SoSciSurvey version of Experiment 1 (SoSciSurvey and jsPsych1). Indexing of Estimate column denotes which transformation was applied to the dependent variable. *** = p < .001; ** = p < .01; * = p < .05

Model	Formula			
SoSciSurvey full model, no outlier correction	log(rt) ~ 1 + task + relatedness + repetition + task:relatedness + (1 + task | subject) + (1 + task + relatedness | picture)			
Fixed effects	Estimate$_{\text{log}}$	Std. Error	*t*-value	*p*-value
Intercept	6.87	0.02	316.99	< .001***
Task	0.3	0.03	10.09	< .001***
Relatedness	0.01	0.01	2.06	.046*
Repetition	0.13	0.01	23.96	< .001***
Task x Relatedness	-0.001	0.01	-0.09	.93

Random effects	Variance	Std. Deviation		
Subjects				
Intercept	0.02	0.13		
Task	0.04	0.19		
Pictures				
Intercept	0.005	0.07		
Task	0.003	0.05		
Relatedness	0.001	0.03		
Residual	0.05	0.22		
Goodness of fit				
Log likelihood	371.4			
Model	Formula			
SoSciSurvey nested, no outlier correction	log(rt) ~ 1 + task/relatedness + repetition + (1 + task | subject) + (1 + task + relatedness$_{\text{naming}}$ | picture)			
Fixed effects	Estimate$_{\text{log}}$	Std. Error	*t*-value	*p*-value
Intercept	6.87	0.02	317.09	< .001***
Task	0.3	0.03	10.08	< .001***
Relatedness$_{\text{BP}}$	0.01	0.01	1.80	.07
Relatedness$_{\text{naming}}$	0.02	0.01	1.50	.14
Repetition	0.12	0.01	23.95	< .001***
Random effects	Variance	Std. Deviation		
Subjects				
Intercept	0.02	0.13		
Task	0.04	0.19		

Pictures		
Intercept	0.005	0.07
Task	0.003	0.05
Relatedness$_{\text{naming}}$	0.002	0.05
Residual	0.05	0.22
Goodness of fit		
Log likelihood	375.7	

jsPsych1. For *jsPsych1*, the Box-Cox procedure suggested transforming the response latency variable by raising to the power of -0.5 (i. e. 1 divided by the square root of the variable). This type of transformation reverses the sign of a model's parameter estimates compared to log-transformed or untransformed data. We therefore transformed using -1 in the nominator (-1/square root of the variable) to maintain the same sign as in the other (log-transformed) models. In the final model, both main effects *task* and *relatedness* were significant, while their interaction was not significant. In the final nested model, the nested effect of *relatedness* was significant in the button press task and marginally significant in the overt naming task. See Table 3 for an overview of the reported models from the preregistered analysis including model formula, coefficients and random effect variance parameters.

Table 3

Table of final models from the preregistered analysis of the jsPsych1 version of Experiment 1 (SoSciSurvey and jsPsych1). Indexing of estimate column denotes which transformation was applied to the dependent variable. *** = p < .001; ** = p < .01; * = p < .05

Model	Formula			
jsPsych1 full model, no outlier correction	-1/sqrt(rt) ~ 1 + task + relatedness + repetition + task:relatedness + (1 + task || subject) + (1 + task + relatedness || picture)			
Fixed effects	Estimate$_{\text{-1/sqrt}}$	Std. Error	*t*-value	*p*-value
Intercept	-0.03	0.0003	119.44	< .001***
Task	0.002	0.0003	6.68	< .001***
Relatedness	0.0002	0.0001	2.23	.031*
Repetition	0.002	0.0001	23.59	< .001***
Task x Relatedness	0.00004	0.0001	0.28	.77
Random effects	Variance	Std. Deviation		
Subjects				
Intercept	0.000002	0.001		
Task	0.000004	0.002		
Pictures				
Intercept	0.0000009	0.001		
Task	0.0000005	0.001		
Relatedness	0.0000002	0.0004		
Residual	0.000008	0.003		
Goodness of fit				
Log likelihood	30889.8			

Model	Formula			
jsPsych1 nested, no outlier correction	-1/sqrt(rt) ~ 1 + task/relatedness + repetition + (1 + task || subject) + (1 + task + relatedness$_{\text{naming}}$ || picture)			
Fixed effects	Estimate$_{\text{-1/sqrt}}$	Std. Error	*t*-value	*p*-value
Intercept	-0.03	0.0003	119.53	< .001***
Task	0.002	0.0003	6.68	< .001***
Relatedness$_{\text{BP}}$	0.0002	0.0001	2.07	.038*
Relatedness$_{\text{naming}}$	0.0002	0.0001	1.77	.08
Repetition	0.002	0.0001	23.59	< .001***
Random effects	Variance	Std. Deviation		
Subjects				
Intercept	0.000002	0.001		
Task	0.000004	0.002		
Pictures				
Intercept	0.0000009	0.001		
Task	0.0000005	0.001		
Relatedness$_{\text{naming}}$	0.0000003	0.001		
Residual	0.000008	0.003		
Goodness of fit				
Log likelihood	30888.8			

Exploratory analyses. While in both versions of the experiment the *relatedness* effect was significant in the models containing the main effect of *relatedness* across tasks and did not interact with the factor *task*, *relatedness* did not reach significance when examining the effect separately for the overt naming task. It is plausible to assume that the overt naming data from an online environment would suffer from an increased level of noisiness and this is also evident in the longer tails of the response time data when comparing the single trial data from Experiment 1 to the data from Abdel Rahman and Aristei (2010), see Figure 4. Therefore, we performed outlier correction employing an approach specifically tailored to LMMs suggested by Baayen and Milin (2010). This approach relies on model criticism after model fitting rather than a priori screening for extreme values, by removing those data points with absolute standardized residuals that exceed 2.5 standard deviations. Baayen and Milin demonstrated that this approach proves more conservative (i. e. excludes fewer data points) compared to more traditional approaches to outlier correction. For the nested models, this led to the exclusion of 2.39 % of trials for the prescreened *SoSciSurvey* data and 1.75 % of trials for the prescreened *jsPsych1* data.

Refitting the final nested model with the outlier corrected *SoSciSurvey* data, the nested effect of *relatedness* was significant in the button press task but not significant in the overt naming task. In the refitted final model of the outlier corrected *jsPsych1* data, the nested effect of *relatedness* was significant in the button press task and in the overt naming task. See Table 4 for an overview of the reported models with outlier corrected data including model formula, coefficients and random effect variance parameters.

Table 4

Table of final models from the exploratory analysis of outlier corrected data from Experiment 1 (SoSciSurvey and jsPsych1). Indexing of Estimate column denotes which transformation was applied to the dependent variable. *** = p < .001; ** = p < .01; * = p < .05

Model	Formula
SoSciSurvey nested, with outlier correction	log(rt) ~ 1 + task/relatedness + repetition + (1 + task | subject) + (1 + task + relatedness$_{\text{naming}}$ | picture)

Fixed effects	Estimate$_{\text{log}}$	Std. Error	*t*-value	*p*-value
Intercept	6.86	0.02	317.58	< .001***
Task	0.3	0.03	10.29	< .001***
Relatedness$_{\text{BP}}$	0.01	0.01	2.17	.03*
Relatedness$_{\text{naming}}$	0.01	0.01	1.1	.28
Repetition	0.13	0.01	27.15	< .001***
Random effects	Variance	Std. Deviation		
Subjects				
Intercept	0.02	0.13		
Task	0.04	0.2		
Pictures				
Intercept	0.005	0.07		
Task	0.003	0.05		
Relatedness$_{\text{naming}}$	0.002	0.05		
Residual	0.04	0.2		
Goodness of fit				
Log likelihood	1140.5			
Model	Formula			
midrule() *jsPsych1* nested, with outlier correction	-1/sqrt(rt) ~ 1 + task/relatedness + repetition + (1 + task || subject) + (1 + task + relatedness$_{\text{naming}}$ || picture)			
Fixed effects	Estimate$_{\text{-1/sqrt}}$	Std. Error	*t*-value	*p*-value
Intercept	-0.03	0.0003	118.11	< .001***
Task	0.002	0.0003	6.86	< .001***

Relatedness$_{\text{BP}}$	0.0002	0.0001	2.17	.03*
Relatedness$_{\text{naming}}$	0.0003	0.0001	2.07	.045*
Repetition	0.002	0.0001	24.36	< .001***
Random effects	Variance	Std. Deviation		
Subjects				
Intercept	0.000002	0.001		
Task	0.000004	0.002		
Pictures				
Intercept	0.0000009	0.001		
Task	0.0000006	0.001		
Relatedness$_{\text{naming}}$	0.0000004	0.001		
Residual	0.000007	0.003		
Goodness of fit				
Log likelihood	30854.8			

Discussion

The results of the two versions of the first experiment were promising regarding the demonstration of the semantic interference effect in an online setting. For both versions we found a significant effect of *relatedness*, with slower response latencies when a picture was accompanied by a semantically related written distractor word compared to an unrelated distractor, replicating the classic picture word interference effect observed in lab settings. As we did not find an interaction of the effect of *relatedness* with the factor *task*, this effect appears to be independent of the task.

As the particular focus of the current study was demonstrating that voice onset latencies could be collected in online settings, we examined the two tasks separately to investigate the effect of *relatedness* specifically in the overt naming task. When looking at the effect in both tasks separately via nested models, we found that the effect did not reach significance in the overt naming task when we ran the preregistered analyses without any outlier correction. When we applied outlier correction by using model criticism and ran the models on the corrected data, the *relatedness* effect in the overt naming task reached significance in the *jsPsych1* version, but not in the *SoSciSurvey* version (where the t value of the estimate actually decreased). An additional model of the overt naming data from both experimental platforms did not yield a significant interaction of the factors platform (*SoSciSurvey* vs. *jsPsych*) and *relatedness* ($b = 0.001$, $t = 0.153$, $p = 0.878$). The absence of this interaction indicates that it is not possible to conclude that the relatedness effect is actually stronger in one platform compared to the other. Even so, the significance of the effect depended on an outlier correction which, while specifically tailored to single trial data in the context of LMMs, we had not planned and therefore not preregistered. It is very likely and plausible that response latency measurements collected via the audio recordings employed in our online experiments suffer from increased random error, i. e. noise, which would decrease the power to find effects.

Presumably, one of the factors contributing to this increased noisiness of the data is the technical implementation of the audio recordings and the reliability of the recordings' timing relative to stimulus onset. This is reflected, for example, in the issue of the variability of the audio file lengths. Only very few of the participants' audio files were exactly of the anticipated length of 3000 ms. Presumably deviations from this file length are due to factors like audio sampling rate, technical variability between the users' machines and fluctuation in internet connection quality. It is beyond the scope and goal of this study to address the technical details of the recording process and to solve the problems underlying the variable file lengths. As long as the file lengths for a single participant were homogeneous we expected the recording process for that participant to be reliable enough to determine reliable speech onsets. Importantly, in the majority of datasets within the *jsPsych1* version the file lengths were homogenuous *within* any one participant, and only a few participants (10 out of 59 participants) showed considerable variability of file lengths (i. e. in more than 20 % of files). We excluded and replaced these ten participants as we could not rule out that in shorter audio files the recording started later than programmed and thus we might not be able to infer the correct voice onset timing.

In the *SoSciSurvey* audio files, the issue was a lot more pervasive. In contrast to the *jsPsych1* data set, 31 of 48 participants included in the final dataset of the *SoSciSurvey* version showed within-subject variability of file-lengths in more than 20 % of files. The issue was present more frequently than absent, which meant that excluding and replacing these participants would have further escalated the already large number of participants required to be collected before reaching the target sample size. The implementation of audio recordings employed in *SoSciSurvey* would therefore seem to be more susceptible to the variability within any one participant's technical set-up. If the variability of the audio file-length is an indication of the reliability of the timing of the audio recording within the experiment, the measurements in *jsPsych* are more reliable.

An additional issue concerning the reliability of the timing of audio recordings in both platforms is the overall difference between the response latencies in the overt naming task of the two versions: voice onset latencies were quicker in *SoSciSurvey* compared to *jsPsych1* ($M = 867$ ms across *relatedness* for *SoSciSurvey* vs. $M = 996$ ms for *jsPsych1*). We cannot account for this difference as the two versions of the experiment were nearly identical and therefore interpret it as a technical issue in the audio recording implementation of the *SoSciSurvey* version. This would align with the finding that the *relatedness* effect in the jsPsych version could also be found when looking at the overt naming task separately in the nested model after exclusion of outliers with a model criticism procedure.

Nevertheless, even the *jsPsych1* version had its difficulties and the effect in the overt naming task was only marginally significant in our preregistered analysis. To assess if the effect in this version was a stable finding or a spurious result, we decided to conduct a second experiment using the same implementation with jsPsych as experiment builder and JATOS as server with several minor adjustments.

One apparent problem with both versions of the first experiment was the high number of data sets which had to be excluded according to our preregistered inclusion criteria. The majority of the excluded participants made too many errors in the button press task, often erroneously classifying the written distractor words' last letter instead of the targets'. Adjusting the instructions in the follow-up experiment to be clearer with respect to the button press task, and providing examples of correct responses to practice trials should improve the error rates and therefore make data collection more efficient. In another adjustment to address the problem of participant exclusion rates, we decided to recruit the participants for the second experiment via the institute's participant pool instead of Prolific, as these participants may be more accostumed to reaction time experiments similar to the current study.

Furthermore, collecting a second dataset using the jsPsych experiment builder would also allow to pool both datasets in a separate analysis, thereby increasing the power to find an effect.

Experiment 2

Experiment 2 was separately preregistered on AsPredicted.com (AsPredicted#: 49281, available at https://aspredicted.org/blind.php?x=9q2yf3).

Methods

The second experiment (*jsPsych2*) was identical to *jsPsych1* with the exception of a few changes.

Participants. In total, 69 native German speakers were recruited using the institute's participant pool Psychologischer Experimental Server Adlershof (PESA). They were included in the final sample when meeting all exclusion criteria until the final sample consisted of 48 participants as determined in the preregistration (36 females, 18 - 35 years, M_{age} = 23.69,

$SD_{age} = 4.99$). Participants provided informed consent to their participation in the study. The study was conducted based on the principles expressed in the Declaration of Helsinki and was approved by the local Ethics Committee. Participants received course credit or a monetary compensation.

Procedure. To increase the efficiency of the data collection and to decrease the high error rates in the first two versions of Experiment 1, an explicit instruction to ignore the written distractor words was included. In addition, each practice trial was followed by an example of what the correct response should have been.

Data exclusion. Participants were excluded and replaced based on the same criteria as in Experiment 1, see Figure 1 for an overview. Similar to the *jsPsych* version of Experiment 1, the issue of irregular file lengths only occurred in a few participants (6 of 69), which were excluded and replaced. In the data set of the final 48 participants, trials were excluded if no response was given, participants made an erroneous response, or if participants responded prematurely. See Table 1 for an overview of the data exclusion of single trials. For Experiment 2, outlier correction via model criticism was applied from the beginning and the models reported in Results were fitted to the outlier corrected data. A further 1.67 % of trials were excluded following the model criticism procedure.

Results

Preregistered analysis. In the button press task, the mean response latency was 1188 ms (SE = 7 ms) in the related condition and 1168 ms (SE = 7 ms) in the unrelated condition. In the overt naming task, the mean response latency was 978 ms (SE = 6 ms) in the related condition and 960 ms (SE = 6 ms) in the unrelated condition. See Figure 2 for a depiction of the impact of *task* and *relatedness* on response latencies for all three versions of the experiment and see Figure 4 for raincloud plots of the single trial data and their distribution by task and relatedness for all experiments as well as for the previous study by Abdel Rahman and Aristei (2010).

The Box-Cox procedure suggested the same transformation as for the *jsPsych1* data: -1 divided by the square root of the response latency variable. In the final model of the outlier corrected data, both main effects *task* and *relatedness* were significant, while there was no significant interaction. In the final nested model, the nested effect of *relatedness* was significant in the button press task and in the overt naming task. This indicates that it takes longer to classifiy and name a target picture if it is presented together with a semantically related distractor. See Table 5 for an overview of the reported models of Experiment 2 including model formula, coefficients and random effect variance parameters.

Table 5

Table of final models from the preregistered analysis of Experiment 2 (jsPsych2). Indexing of Estimate column denotes which transformation was applied to the dependent variable.

*** = p < .001; ** = p < .01; * = p < .05

Model	Formula			
jsPsych2 full model, with outlier correction	-1/sqrt(rt) ~ 1 + task + relatedness + repetition + task:relatedness + (1 + task || subject) + (1 + task + relatedness || picture)			
Fixed effects	Estimate$_{\text{-1/sqrt}}$	Std. Error	*t*-value	*p*-value
Intercept	-0.03	0.0003	118.44	< .001***
Task	0.003	0.0003	10.89	< .001***
Relatedness	0.0002	0.0001	2.73	< .01**
Repetition	0.002	0.0001	28.78	< .001***
Task x Relatedness	0.00003	0.0001	0.24	.81

Random effects	Variance	Std. Deviation		
Subjects				
Intercept	0.000002	0.001		
Task	0.000003	0.002		
Pictures				
Intercept	0.0000009	0.001		
Task	0.0000004	0.001		
Relatedness	0.0000002	0.0004		
Residual	0.000007	0.003		
Goodness of fit				
Log likelihood	31715.2			
Model	Formula			
jsPsych2 nested, with outlier correction	-1/sqrt(rt) ~ 1 + task/relatedness + repetition + (1 + task || subject) + (1 + task + relatedness$_{\text{naming}}$ || picture)			
Fixed effects	Estimate$_{\text{-1/sqrt}}$	Std. Error	*p*-value	*p*-value
Intercept	-0.03	0.0003	118.47	< .001***
Task	0.003	0.0003	10.84	< .001***
Relatedness$_{\text{BP}}$	0.0002	0.0001	2.67	.008**
Relatedness$_{\text{naming}}$	0.0003	0.0001	2.18	.035*
Repetition	0.002	0.0001	29.02	< .001***
Random effects	Variance	Std. Deviation		
Subjects				
Intercept	0.000002	0.001		
Task	0.000003	0.002		

Pictures		
Intercept	0.0000009	0.001
Task	0.0000004	0.001
Relatedness$_{\text{naming}}$	0.0000003	0.001
Residual	0.000006	0.003
Goodness of fit		
Log likelihood	31702.3	

Discussion

The results of Experiment2 (*jsPsych2*) confirmed the results from the *jsPsych1* version of Experiment 1. The effect of relatedness was significant in the full model as well as in both tasks separately in the nested model, replicating the semantic interference effect in the PWI in general, and the findings from Abdel Rahman and Aristei (2010) in particular. The implementation of audio recordings in the online environment offered by jsPsych and JATOS seems to be able to provide stable measurements of verbal response latencies.

The adjustments to the procedure in *jsPsych2* also improved the efficiency of the online data collection. Whereas in *jsPsych1,* datasets from a total of 108 participants needed to be collected to reach the desired sample size of 48, only 69 participants were required in experiment *jsPsych2* to reach the same goal. While the main reason for exclusion was still participants' error rate in the button press task, the number of participants with an error rate above 20 % decreased from 43 participants in *jsPsych1* to 13 participants in *jsPsych2.* Furthermore, most of these 13 participants had an error rate only slightly above our predefined threshold indicating that they did not misunderstand the task and classified the distractor instead of the target, which had been the case for *jsPsych1.*

Pooled analysis and post-hoc analyses of power

To increase the power of the analysis, the data from *jsPsych1* and *jsPsych2* was pooled. The model criticism procedure for the pooled data resulted in the exclusion of an additional 1.71 % of trials, compared to the pooled data of the two experiments without any outlier correction.

The Box-Cox procedure suggested the same transformation as for the *jsPsych1* and the *jsPsych2* data: -1 divided by the square root of the response latency variable. In the final model, both main effects of *task* and *relatedness* were significant, in the absence of an interaction. In the final nested model, the nested main effect of *relatedness* was significant in the button press task and in the overt naming task. See Table 6 for an overview of the reported models from the pooled analysis including model formula, coefficients and random effect variance parameters.

Table 6

Table of final models from the preregistered analysis of the pooled analysis (jsPsych1 + jsPsych2). Indexing of Estimate column denotes which transformation was applied to the dependent variable. *** = p < .001; ** = p < .01: * = p < .05

Model	Formula			
Pooled *(jsPsych1 + jsPsych2)* full model, with outlier correction	-1/sqrt(rt) ~ 1 + task + relatedness + repetition + task:relatedness + (1 + task || subject) + (1 + task + relatedness || picture) + (0 + task || experiment)			
Fixed effects	Estimate$_{\text{-1/sqrt}}$	Std. Error	*t*-value	*p*-value
Intercept	-0.03	0.0002	144.97	< .001***
Task	0.003	0.0003	7.83	.007**
Relatedness	0.0002	0.0001	2.45	.008**
Repetition	0.002	0.00004	35.24	< .001***
Task x Relatedness	0.00006	0.0001	0.75	.49

Random effects	Variance	Std. Deviation
Subjects		
Intercept	0.000002	0.001
Task	0.000004	0.002
Pictures		
Intercept	0.000001	0.001
Task	0.0000005	0.001
Relatedness	0.0000002	0.001
Experiment		
Task	0.0000001	0.0003
Residual	0.000007	0.003
Goodness of fit		
Log likelihood	62624.9	

<table>
<tr><th>Model</th><th>Formula</th></tr>
<tr><td><i>Pooled (jsPsych1 + jsPsych2)</i> nested, with outlier correction</td><td>-1/sqrt(rt) ~ 1 + task/relatedness + repetition + (1 + task || subject) + (1 + task + relatedness_{naming} || picture)</td></tr>
</table>

Fixed effects	Estimate$_{\text{-1/sqrt}}$	Std. Error	*t*-value	*p*-value
Intercept	-0.03	0.0003	145.03	< .001***
Task	0.003	0.0003	7.84	.01*
Relatedness$_{\text{BP}}$	0.0002	0.0001	2.22	.008**
Relatedness$_{\text{naming}}$	0.0003	0.0001	2.1	.02*
Repetition	0.002	0.0001	35.24	< .001***

Random effects	Variance	Std. Deviation
Subjects		
Intercept	0.000002	0.001
Task	0.000004	0.002
Pictures		
Intercept	0.000001	0.001
Task	0.0000005	0.001
$\text{Relatedness}_{\text{BP}}$	0.0000001	0.0002
$\text{Relatedness}_{\text{naming}}$	0.0000003	0.001
Experiment		
Task	0.0000001	0.0003
Residual	0.000007	0.003
Goodness of fit		
Log likelihood	62624.9	

As expected, the effect of *relatedness* in both tasks was more stable when pooling the data from two experiments and thus increasing the power of the analysis. To determine to what extent this pooled analysis might have been “overpowered” and to find a balance between sufficient power on the one hand and sensible sample sizes on the other hand, we performed a post-hoc power analysis with the package simr (Green & MacLeod, 2016) using the parameter estimates derived from a separate model of only the overt naming task from the pooled data.[9]

For the post-hoc power analysis we calculated a power curve, which relies on 1000 simulated data sets based on parameters from our pooled data set. These simulated data sets were then

[9] Researchers wishing to analyse data of future experiments with ANOVAs can find the results of F_1/F_2 within-subjects ANOVAs including effect sizes in the online Supplementary Material B to use for a priori power analyses (https://osf.io/uh5vr/?view_only=229679aa33604aa2a5cb400eab62099).

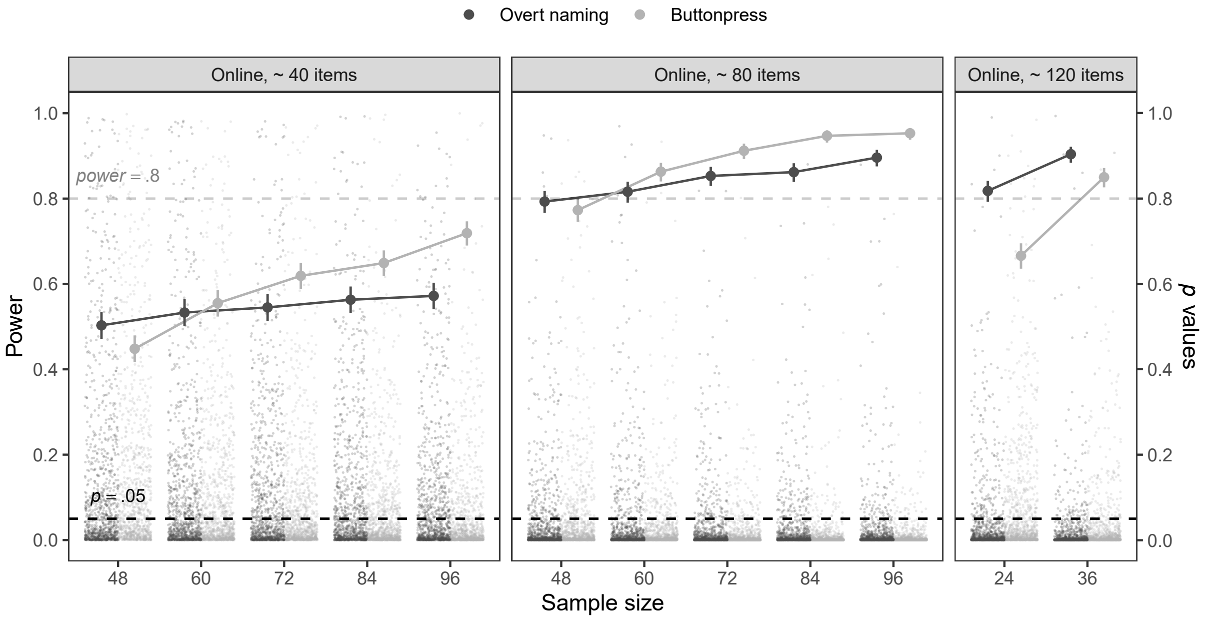

Figure 5. Results of the post-hoc power simulations for the fixed effect of *relatedness* in both tasks based on estimates from the pooled analysis with an increase in both sample size (on the x-axis) and number of items (different panels). The big dots represent power plotted by different sample sizes. For each sample size the number of simulations to estimate power was n = 1000. The small dots represent the resulting *p*-values for each of the 1000 simulations. Increases in power result from higher proportions of runs with *p*-values below the threshold of $p = .05$. The dashed grey line represents the threshold for reaching a power of 80 %.

analyzed with the proportion of significant results relative to all simulations indicating the respective power (Kumle, Võ, & Draschkow, 2021). Figure 5 displays the estimated power for increasing sample sizes and increasing number of trials.

The observed power to find a significant effect of *relatedness* increased with growing sample sizes as expected. At a sample size of 96 (i. e. the actual sample size of the pooled analysis) and 40 trials as in this experiment the simulated power is 72 % for the button press task and 60 % for the overt naming task. These values are notably lower than the 80 % power with a sample size of 36 participants that we determined via an a priori power analysis (see Methods section of Experiment 1). A possible reason for these differences is the fact that for the a priori analysis we relied on estimates from a lab-based study and only included random intercepts, whereas the model estimates used in the post-hoc analysis included a random slope parameter for item as well as a correlation parameter of the random intercept and slope by item, as determined empirically by our model selection process. This resulted in an increase in the number of parameters between the models from four in the a-priori analysis to six in the post-hoc analysis. Matuschek, Kliegl, Vasishth, Baayen, and Bates (2017) point out that fitting more complex models with more random effect parameters comes at a cost of power. Furthermore, it is notable that the increase in power with increasing sample sizes is not very large, especially for the overt naming task. However, increasing the number of items seems to be more beneficial and important than increasing the observations within each subject in order to reach an estimated power of 80 %. Indeed, when running simulations increasing both sample size and the number of items, we see that the power strongly increases from 40 to 80 items. For example, a study with 60 participants and 80 items or with 36 participants and 120 trials would yield an expected power of 80 %. Thus, under some conditions future experiments may profit more from an increase in items rather than an increase in subjects. However, see our recommendations in the General Discussion for possible drawbacks to this approach.

General Discussion

In the present study we introduced three online implementations of the PWI task and replicated the well-known semantic interference effect. To the best of our knowledge, this is the first time stimulus-locked voice recordings from an online experiment have been used to succesfully measure voice onset latencies as a dependent variable, providing a proof of concept that language production experiments relying on overt naming can be moved online.

In our implementations of an online PWI task, we presented pictures with visually superimposed distractor words that were either semantically related or unrelated to the target picture. Participants were asked to name the picture (overt naming) or, as a control task, classify the last letter of the picture name as a vowel or consonant (button press). Our goal was to find the typically observed semantic interference effect with longer response latencies if a distractor word and a picture are semantically related (vs. unrelated). Given the poor audiovisual synchrony reported for different web experiment builders, browsers, and hardware configurations (Bridges et al., 2020) in combination with the small size of the semantic interference effect of around 20 ms, this was not a trivial endeavor. Despite these challenges we replicated the semantic interference effect in both tasks, classification and overt articulation in three pre-registered experiments, and the effect closely resembles in size the effect reported by Bürki et al. (2020) in their recent metastudy on semantic interference in the PWI task. We conclude that running language production experiments online is feasible.

Data quality in our implementation of online language production experiments

Data loss. Comparing the amount of data loss over the course of the three experiments with studies run in the lab, it is evident that a higher number of participants had to be tested in order to reach our predefined goal of collecting 48 valid data sets. In the first two runs of the study more than double the number of data had to be collected in order to obtain at least 80 % of trials per task and participant for analysis. The reasons for this high loss of data sets are two-fold.

One source of error are the participants themselves. We found that many participants did not read the instructions carefully enough and hence had a high error rate when classifying the last letters. This was especially pronounced in the first experiment where many participants classified the distractor word instead of the target. However, this problem was minimized in Experiment *jsPsych2* where we used a participant pool that might be more accustomed to lab settings and by giving more explicit instructions as well as providing performance feedback by giving the correct response in the practice trials. Although the number of participants with an error rate above 20 % was still non-negligible, the number of participants always classifying the distractor word was reduced substantially by these measures.

The second reason for loss of data sets can be subsumed under technical problems. We encountered cases with empty audio files, differing file lengths within participants, noise on audio files, as well as files of poor audio quality. Note that there were no empty audio files in Experiment 2, indicating that empty files might be a result from non-compliant participants muting their microphones (we presume our participants were more compliant in Experiment 2).

Noisiness of data. In our experience, data collected online was noisier than data collected in the lab with longer tails in the distribution of response time data compared to the lab experiment by Abdel Rahman and Aristei (2010) and longer overall response times. This has also been found by other groups running online language production studies recently (Fairs & Strijkers, 2021). We deem it likely that the lack of a controlled testing environment when testing online is the reason for these relatively long reaction times. However, as can be seen from Figure 3, we do find classic repetition effects with participants getting faster with repeated stimulus presentations. This underlines our conclusions that online testing is a suitable approach for using language production paradigms relying on the estimation of voice onset latencies.

Still, even though we accounted for the potential greater noisiness of online data a priori by raising the number of participants by a third after running a power estimation for the effect from a previous lab experiment, the interference effect for naming was only marginally significant in Experiment *jsPsych1.* To counter noise in the data we performed an outlier screening by applying model criticism. This improved the quality of the obtained data but has not been necessary to obtain interference effects in lab-based experiments. Furthermore, pooling data from both jsPsych experiments and thus increasing sample size made the interference effect more stable. Therefore, it seems likely that we underestimated the noisiness of online language production data in our first experiment.

Efficiency. In contrast to data collected in the lab using a voice key, the data generated online were not ready for analysis after their collection. First, voice onset latencies were determined by using Chronset (Roux et al., 2017) and these were manually corrected as latencies were not computed reliably in all cases by this software tool. This procedure consumes time and resources, but could maybe be optimized in future.

In summary, contrary to other fields of behavioral psychology, running language production experiments online is not (yet) less resource consuming than research in the lab in order to obtain data sets of sufficiently high quality. We hope that future work might help to reduce these efforts. In the following we provide recommendations based on our experiences that will likely improve the quality of the data.

Recommendations for running online language production experiments

Take an informed decision on which web experiment builder you use. The choice of a web experiment builder may have a strong impact on the quality of the data in a web experiment (Bridges et al., 2020). Based on our data we can strongly recommend using jsPsych (de Leeuw, 2015) for experiments in which audio responses are recorded – either implementing the audio-response-plugin which is provided in a beta version (Gilbert, 2020) or by using our custom script available on OSF (https://osf.io/uh5vr/?view_only=229679aa33604aa2a5cb400eab62099).

We tested different ways to implement online audio recordings using two web experiment builders for online experiments – SoSiSurvey (Leiner, 2019) and jsPsych (de Leeuw, 2015). Overall we encountered less technical comparisons for jsPsych in comparison with the other experiment builder. Less audio files had to be discarded due to low quality and thus data loss was minimized to a substantial degree using jsPsych in combination with a custom audio record plugin. We therefore advocate the use of jsPsych, a non-commerical and open source software library for building web-based experiments with a proven record in a wide range of behavioral experiments. Many experimental tasks can easily be built by using the experimental plugins provided and even with little experience in JavaScript programming there are almost infinite possibilities to fine-tune them to cater to the needs of the experimenter. Furthermore, there is a very active and commited helper community available for questions that might arise during the process of developing the experiment. Presumably, using an audio recording option provided by other web experiment builders (e.g., LabVanced, Gorilla, Finding Five) would have led to similar results. However, no recording options for these experiment builders were available when preparing this study and a thorough examination of their timing reliability is still outstanding.

Custom audio recording implementation in jsPsych. As none of the available web experiment builders offered an audio recording option ensuring that audio recordings would be precisely time-locked to other stimuli, we customized a jsPsych plugin to our needs. With this plugin the experiment proceeds in the following way: When starting the experimental block, participants have to grant the browser access to their microphones once and microphone access remains active over the whole naming part. Within the naming part, the custom audio record plugin enables recording of short audio files of a predefined duration (in our case three seconds) timelocked to the presentation of another stimulus (in our case a picture). The experiment then proceeds as defined in the experiment timeline, e.g., by presenting a fixation cross or the next trial. In the background, the audio files are transferred to the server where the experiment is hosted. We chose to transfer the files to a server using JATOS (Lange et al., 2015), an open source tool for running online studies on your own server. Furthermore, JATOS also offers the option to structure the experiment in different components that are executed one after the other – a feature that we used in order to only have microphone access enabled in the naming part of the experiment but not during the following or preceding button press task. However, the custom plugin should in principle be compatible with any other server that might best serve a researcher's needs when hosting the experiment and saving the data. After the experiment, the audio files and log files can be downloaded and saved. Voice onset latencies can be extracted from the audio files and the onset latencies can be merged with the logfile from the experiment.

Limit noise stemming from technical side. It is essential to limit potential sources of noise stemming from the technical side. Noise can be introduced by variation in participants' choice of browsers, their specific hardware and software. While it is not possible to have full control over the technical equipment of participants in an online experiment, it is possible to account for some of their potential influence. For example, we monitored which browser participants used in order to control whether data collected in one browser might be more or less reliable. We did not find evidence for specific browser related differences in the reaction time data in the jsPsych experiment. However, during pretesting we found that the experiment was not working for Edge and sometimes not for Safari users. Therefore, we advised participants to run the experiment on a personal computer using Firefox or Chrome, thereby also minimizing the potential influence of different devices and browser types.

Furthermore, we strongly suggest to only use within-participant designs, which may reduce variability due to participants' hardware and software setups and can help to minimize the influence of audiovisual synchrony problems on the experimental manipulation.

While future work on web experiment builders might help to reduce these technical problems, we deem it important to thoroughly screen the data gained from a web based language production experiment and in case of doubt, rather exclude participants than keep participants with deficient data. To provide transparency to these decisions we suggest preregistering sample size and data-exclusion criteria.

Limit noise stemming from participants. Samples reached via online testing may not be as cooperative and accustomed to the prerequisites of experiments as the standard lab population. Therefore, it is also essential to limit potential noise stemming from participants. For example, noise can be reduced by thoroughly instructing participants using catch questions to test their understanding of the task (Oppenheimer, Meyvis, & Davidenko, 2009) and by giving ample feedback during practice trials. Furthermore, the pool from which participants are recruited may have an impact on data quality, as evident by the reduced number of data sets that had to be excluded in Experiment 2 for which we recruited participants from our institute's subject pool.

Note that screening procedures and sampling strategies may ensure higher data quality while potentially minimizing the advantage of testing more diverse samples than in the lab.

The chosen selection criteria may induce a sampling bias as e.g., stable internet connections, well-set hardware or the willingness to grant access to a microphone are not equally distributed. Therefore, researchers should be aware of the fact that online testing does not entail more representative samples than typical samples in the lab per se. Researchers should aim for a balance between high data quality and minimal sampling bias when planning online experiments.

Account for increased noise by increasing sample size. We suggest to estimate a sample size based on previous data and to increase the estimated sample size by at least 33 % to account for noise due to online testing. Not all sources of noise can be totally controlled and minimized by the experimenter. Therefore, the amount of data needed to be able to draw sound statistical inferences is likely higher compared to lab settings. Researchers might as well choose to increase the number of trials in their experiment. However, it is advisable to keep online studies short because long, and possibly boring, tasks may lead to an attrition of participants' attention or increase the rate of participants who abort the study (Sauter et al., 2020).

Carefully check quality of data after the experiment. After the experiment, it is important to carefully check the recorded audio files, e.g., by listening to the files to check the audio quality and to control whether participants answered correctly. Additionally, researchers should inspect the files for differing file lengths within any one participant. Furthermore, we suggest to check the estimated voice onsets and to use an outlier screening procedure before analysing the data. Of course the best way to check the quality of your data is to replicate your basic effect first.

Open Issues

Differing file lengths. We lost several data sets before data analysis due to different audio file lengths within single participants. In most cases the audio files stemming from the same participant had the same duration. However, there were cases where we encountered different file durations within the data stemming from the same participant. We thoroughly investigated this issue and did not find any correlation between differing file length and any of the software or hardware configurations we had logged (browser used, operating system used). We were not able to clarify whether recordings started late (which would be fatal to an accurate estimation of response time), or were cut in the end (which would be less problematic). We therefore took a conservative approach and replaced participants for whom differing file lengths occurred in more than 20 % of the audio files. We hope that future research and software development will help to tackle this problem and thus make the online collection of precisely timed naming responses necessary for many language production paradigms more efficient.

Potentially poor audiovisual synchrony. Lack of audiovisual synchrony is a documented challenge for web based experiments (see also Bridges et al., 2020; Reimers & Stewart, 2016). We cannot quantify to what degree the problem of poor audiovisual synchrony, a lag of differing duration between presentation of a picture and start of the audio file creation, existed in our experiment, too. The trial duration itself is predetermined via the specific parameters set while programming the experiment. It is logged in the respective logfiles and has a high reliability, i. e. the actual trial duration corresponds to the predefined trial duration. However, within each trial, several events need to be executed by the browser: a visual stimulus needs to be presented and an audio recording needs to be started. Therefore, audiovisual synchrony can be poor for two reasons. First, there may be delays between when the visual stimulus onset is requested and when it actually appears on screen and second, there may be delays between the recording request and when it actually starts. While there exist technical solutions to minimize the first problem, it is still a task for Java Script developers to minimize the delays between the recording request and its start (Bridges et al., 2020). Without special technical equipment it is not possible to log how long it takes for a request to be executed and thereby to quantify the problem of audiovisual synchrony for each and every participant. One way to potentially quantify the problem of audiovisual synchrony would be to externally monitor participants' screens for the appearance of a stimulus, displaying an audio signal like a beep immediately upon stimulus appearance from an external device which will then be recorded on the audio file. Later, a comparison of the latency difference between requested stimulus onset and audio file offset with the beep signal onset on the audio file in relation to the audio file offset needs to be done. Obviously, this is not possible when running experiments online. Therefore, a careful examination of the relative timing of events within a single trial is beyond the scope of this article while it may in principle be done (Gilbert & Minors, 2020). However, given our replication of a small effect of 20 ms we deem it reasonable that the problem of audiovisual synchrony can be neglected and reaction times can be estimated to a degree of accuracy that was sufficient for our purpose. For the future, we recommend monitoring the developments tackling the issue of audiovisual synchrony. For the time being it should be kept in mind that even though options for voice recording are available and the software allows to define when and for how long an audio file is recorded, this does not necessarily guarantee that the timing is sufficiently well controlled for in order to draw inferences from voice onset latencies.

Future avenues

One of the biggest advantages of conducting language production experiment online is that language production research will become less dependent on available lab space and thereby become more accessible. This will not only be helpful inmidst the Covid19-pandemic with researcher having to move their experiments online. Furthermore, undergraduate students might be able to run their own small studies without using lab space. Participants from bigger and more diverse samples, even in remote areas, can be accessed more easily as long as they have internet access. Hitherto understudied populations could be more readily featured in language production research – making the field less reliant on the classic WEIRD (= western, educated, industrial, rich, democratic) population (Henrich, Heine, & Norenzayan, 2010). Data from more diverse samples will be essential to test the validity of empirical findings in the language production literature.

With this study we provide a hands-on solution for running language production experiments online. With this proof of concept and alongside our suggestions that were derived from our experiences with implementing the experiments, we are confident that also other types of language production experiments, for example semantic blocking or the cumulative naming task, can be implemented online (Fairs & Strijkers, 2021; Stark, van Scherpenberg, Obrig, & Abdel Rahman, 2022). This will help to address many of the hitherto open questions in language production research (e.g., Abdel Rahman & Melinger, 2019; Bürki et al., 2020).

Acknowledgements

Roger Hauber was supported by a grant from the German Research Foundation (DFG) awarded to Anna K. Kuhlen (AK3236/3-1) and Rasha Abdel Rahman (AB277/11-1). Anne Vogt was supported by a doctoral fellowship of the German National Academic Foundation (Studienstiftung des deutschen Volkes) and the Berlin School of Mind and Brain. We are grateful to Guido Kiecker for help with setting up Experiment 1 in SoSciSurvey and to Özgür Kesim for helping us to customize the jsPsych-plugin and deepen our understanding of web programming. Our thanks also go to all the people in the research community who openly share their code and expertise and helped us to build the experiment by answering our questions, especially Josh de Leeuw, Becky Gilbert, Elisa Filevich and Kristian Lange. Special thanks goes to Mara Haslan whose audio record plugin shared in the jsPsych Google Group formed the base for the audio recording option we implemented in jsPsych. We thank Cornelia van Scherpenberg for help with implementing the correction procedure for voice onset estimations, Clara Niemann for supporting data collection of Experiment 2, Nora Holtz, Annika Just and Clara Niemann for help with preprocessing the data, Nora Holtz and Alexander Enge for help with conducting the post-hoc power estimations. All authors developed the study design, discussed the results, and wrote the manuscript. Anne Vogt implemented the experiment in jsPsych, both Anne Vogt and Roger Hauber performed data collection and analyses.

4 Discussion

The findings from this dissertation advance our understanding of semantic processing in language production. The evidence adds to the growing research on hitherto neglected semantic relations in language production research, namely the role of experientially grounded and embodied meaning aspects. I show that reactivations of experiential traces of space influence our choice of words when speaking. Thereby this work contributes to the long line of research on experientially grounded language processing but with a focus on language production. Moreover, the use of a novel paradigm allowed to analyse the content of the output of lexical selection in contrast to previous work, which investigated mainly the time course of lexical selection. Furthermore, this work highlights the importance of the role of interindividual differences in language production. Both, the importance of embodied meaning aspects in language production as well as the necessity to take interindividual differences into account, add to the strong need of developping better models of semantics in language production.

I will adress experientially grounded meaning aspects in language production in detail in the following Section 4.1, before focusing on the role of interindividual differences in language production in Section 4.2. Following from this evidence, I discuss consequences for models of semantic processing in language production in Section 4.3. In Section 4.4, recommendations are given for running timing-sensitive picture naming experiments online which will enable researchers in the field to test larger, more diverse and less accessible populations in the future, before concluding the dissertation with a brief summary and outlook in Section 4.5.

4.1 The role of experientially grounded meaning and its reactivation during language production

Given the enormous importance of spatial thinking for human cognition (see e.g., Bellmund, Gärdenfors, Moser, & Doeller, 2018), it is perhaps not surprising that spatial meaning aspects get reactivated during semantic processing. As outlined in section 1.1.3, there is indeed much evidence that participants implicitly process the vertical spatial locations implied by words like *sun*, *skyscraper* or *soil* when they read or hear those words (e.g., Bergen et al., 2007; Dudschig et al., 2012; Estes et al., 2008; Ostarek et al., 2019; Ostarek & Vigliocco, 2017; Richardson et al., 2003; Verges & Duffy, 2009). However, evidence from language production was missing so far. Therefore, I investigated whether reactivating the experientially grounded meaning aspect of LOCATION IN SPACE influences the outcome of lexical selection in a language production task.

A sentence completion paradigm was used to examine whether the spatial locations of the referents of the chosen sentence completions would be influenced by spatial manipulations. In Study 1, the visual appearance of sentence fragments was manipulated to the end that sentences seemed to move upwards or downwards on screen. Building on results from the language comprehension literature (Dudschig et al., 2013; Kaschak et al., 2005; Meteyard et al., 2008) it was assumed that the location of the produced words' referents can be predicted by the presentation direction of sentences. In Study 2, the bodily involvement of participants was enhanced. While sentences were presented auditorily, participants had to move their head upwards or downwards with their eyes closed before producing a suitable noun to complete the sentences. Again, it was assumed that the spatial properties of the produced words would be influenced by the head movement direction. Indeed, in a similar study participants recalled more words which were spatially compatible with their head position (Lachmair, Ruiz Fernández, et al., 2016).

Whereas previous work on embodied meaning relations in language production focused on differences in the time course of naming pictures, we investigated the outcome of the lexical selection process proper, i.e. we explored WHAT people say and not WHEN they say it. To this end, all produced words were rated according to the spatial locations of their referents and these spatial properties then served as dependent variable.

4.1.1 Reactivating experiential traces via movements (but not through vision)

We found that reactivating embodied meaning aspects can directly influence which words we choose when speaking. In line with the hypothesis, the head movement direction had a significant impact on the lexical choices of participants in Study 2. Participants chose sentence endings whose referents were more in line with the head movement direction. Contrary to the expectations, no such effect was found for the visual manipulation in Study 1 where sentences had been presented in a downward or upward moving way. The hypothesis for Study 1 had been based on evidence from the language comprehension domain. There, it was shown that visual stimulation can lead to the reactivation of experiential traces of space which then influence language comprehension (Dudschig et al., 2013; Kaschak et al., 2005; Meteyard et al., 2008). However, in an anagram solving task, it was shown that the experimental manipulation needs to bear a significant amount of meaning in order to reactivate experiential traces of space, which subsequently influence semantic processing (Berndt et al., 2018). The anagram solving task is conceptually close to lexical selection and therefore might show that a stronger experiential reactivation is needed to influence lexical selection processes. Probably, the experiential reactivation of vertical space was not sufficient in Study 1 because it was not meaningful enough. Participants did not link the vertical visual manipulation in Study 1 to the content of the sentences and therefore did not use it as a cue for lexical selection. In contrast, the greater involvement of the body in Study 2 enhanced the meaningfulness of the experimental manipulation and the movement itself was incorporated in the process of lexical selection.

As a mechanism for the effect of head movement in Experiment 2, I propose that the movement reactivates experiential traces of space, which follows from embodied language processing theory (see e.g., Zwaan & Madden, 2005). The concurrent sentence completion task leads to the activation of a lexical cohort of potentially suitable sentence endings. Words sharing the experiential trace of space, which is already activated by the movement, receive higher activation and are therefore more likely to be produced. Note that we asked participants after the experiment whether they correctly guessed the purpose of the body movements. In accordance with our preregistered exclusion criteria, across both experiments 9 participants had to be excluded from analysis because they had inferred the idea behind the experimental manipulation or indicated that they deliberatively used the body posture to come up with suitable words. Thus, we only analysed data from participants who did not use their body posture deliberatively and consciously as a cue to find suitable sentence endings.

Investigating language production and embodiment also provides a possibility to draw conclusions on the causality of embodied meaning effects: the head movements seem to have been causally relevant for the lexical choices as we manipulated body posture and subsequently measured a linguistic outcome. Thereby, we addressed one of the challenges faced by embodied language processing researchers, namely probing causality (Ostarek & Bottini, 2020; Ostarek & Huettig, 2019). In contrast, the vertical visual manipulation of word position in Study 1 was not sufficient to influence language production causally.

4.1.2 Reactivating experiential traces embedded in language

Additionally, we found that experiential traces of space embedded in language influence lexical selection. In Study 1 and Study 2 the wider spatial context of the situations described in the sentence fragments influenced lexical choices. For example, participants more likely completed a sentence like <*You are at the beach and you see ...*> with a noun like *shell* compared to a noun like *gull*. The described beach situation is rather down in vertical space, and this made it more likely that participants produced words sharing spatial features with the situation – in this case *shell*, which could presumably lie on the ground.

The cognitive mechanism by which experientially grounded information embedded in language itself is (re-)activated during concept use would be explainable in the same ways as the reactivation of experiential traces via head movements: the words in the sentence fragments reactivate experiential traces, which are connected to previous encounters with these words and their referents. Among these experiential traces may be experiential traces of space. As a consequence, from the cohort of possible nouns to complete the sentence those which share experiential traces with the words in the sentences will be more likely selected.

Note that there was no interaction between the effect of sentence spatial locations and head movements in Study 2. Thus, even though the mechanism leading to the observed effects might be similar, the two ways of reactivating experiential traces seem to target different meaning aspects.

This idea has been captured in hybrid theories of semantics. According to these, our experience with language can be seen as just another way of direct experience from which we acquire semantic knowledge (see Unger & Fisher, 2021, for a review, see e.g., Günther et al., 2020, for empirical evidence). Furthermore, the semantic relations embedded in text often reflect real-world contingencies and thus mirror the world which we directly perceive and experience (Günther et al., 2019). As a result, hybrid models of semantics treat linguistic and perceptually based meanings as highly related, varying in the degree to which they are conceived of as separable (Banks et al., 2021), or essentially inseparable meaning aspects (Davis & Yee, 2021). Therefore, and as an extension to Firth's original proposal, we know the meaning of a word by the linguistic and perceptual company it keeps (Louwerse, 2018).

4.1.3 Alternative explanations

The finding that the spatial features of the produced nouns can be predicted by the spatial properties of the presented sentences raises the question whether a word like *shell* might be more predictable in the context of a sentence about a beach situation. Probably *shell*, which is sharing the spatial features with the sentence noun, is semantically closer to the word *beach* than the word *gull*, which is not sharing the spatial features with the sentence noun to the same degree.

In this case, it would be questionable whether an explanation in terms of shared experiential features of space between the noun describing the situation in the sentence and the produced noun is indeed the most probable explanation.

We tried to tackle this issue by analysing how many different words were uttered for each sentence fragment. Remember, that we had intended to use non-constraining sentence fragments for the study. Post-hoc analysis corroborated that we had succeeded in doing so: In Experiment 2, participants produced between 17 - 42 different words for the different sentence fragments (the maximal number would have been 72, equalling the number of participants) and cloze values ranged from 0.07 - 0.43, which is not considered high (Block & Baldwin, 2010). Thus, we found a substantial degree of variation which words participants chose to complete the sentences. Furthermore, when we only investigated those cases where the predictability was lowest, i.e. 36 or more different words were produced across all participants, the effect of sentences' spatial locations on the spatial properties of the produced nouns persisted. Therefore, we deem it unlikely that predictability led to the observed effect of spatial characteristics of the described sentences influencing the spatial characteristics of the produced words.

In a similar vein, we tried to rule out that participants merely produced semantically close words, which also happen to share the spatial connotations of the described situations and that the effect could therefore be explained by semantic similarity. To this end, we included a distributional measure of semantic similarity in our analysis and computed the semantic similarity between the noun in the stimulus sentence and the produced noun (Günther et al., 2015). Then we explored in how far this influenced the effect of the spatial characteristics of the sentences on the spatial characteristics of the produced words. In both, Study 1 and Study 2, the effect of sentence spatial characteristics on spatial features of the produced nouns was reliably observable across the range of similarity values obtained, even though it got stronger the more similar the produced noun was to the sentence noun. In other words, even in cases of negligent semantic similarity between a sentence and a produced noun we still observed that the chosen nouns were aligned with the broader spatial context of the described situations.

While we tried to rule out that the findings can be explained by predictability and distributional semantic similarity, it should be noted that the notion of predictability is empirically underspecified with regard to language production. Embodied meaning aspects may be among the factors contributing to making a word more likely to be uttered in a language production context. Specifically, this has been shown by the body movement manipulation in Study 2. Furthermore, linguistic distributional information like the frequency of certain phrases may contribute, too (Janssen & Barber, 2012; Shao, van Paridon, Poletiek, & Meyer, 2019). However, which (other) factors contribute to determining the selection of one word over the other remains to be further explored empirically in order to specify what makes a word predictable.

4.2 Interindividual differences

An important tenet from theories of embodiment or grounded cognition is the uniqueness of experiences. As a consequence, interindividual differences in the ways we gain and process experiences may differentially impact the formation of concepts (e.g., Medin et al., 2006; Tanaka & Taylor, 1991) and thereby also language processing.

4.2.1 Interoceptive Sensibility

Importantly, people differ in the sensitivity to perceive body processes and bodily signals, showing different degrees of interoceptive sensibility (Mehling et al., 2012). Due to the strong link between interoception and the body, it follows that embodied cognition and thus also language processing could be influenced by interindividual differences in the sensitivity to signals from inside our bodies (Häfner, 2013).

In Study 2, we investigated whether interindividual differences in interoceptive sensibility lead to differential outcomes in a paradigm where sensitivity to the body could be crucial for the hypothesized effect of body movement on lexical choices. Indeed, we found that the effect of the head movement manipulation on the spatial properties of the produced words was stronger for participants with greater ability to sustain and control attention to body sensations (measured with the *attention regulation* subscale from the MAIA questionnaire, Mehling et al., 2012). Presumably, with higher ability to regulate the attention towards bodily signals participants are more prone to integrate their bodily state into their lexical processing. Participants' awareness of uncomfortable, comfortable and neutral body sensations as indexed by the *noticing* subscale from the same questionnaire did not moderate the effect of head movements on lexical choices.

Recently, similar results have been obtained by investigating word production in Parkinson patients receiving deep brain stimulation (DBS) of the subthalamic nucleus. When DBS was off, patients produced less verbs related to movement of their own body compared to a control group. No such difference was apparent when DBS was on (Klostermann et al., 2022). Thus, the current state of the motor system of Parkinson patients can lead to differences in semantic processing in language production. Together with the results from Study 2, this demonstrates that we can gain interesting insights by incorporating interindividual differences as factors which might explain some of the hitherto unexplained variance in language production studies. To date, research on interindividual factors and their role in semantic processing during language production has been very sparse (but see e.g., Hughes & Schnur, 2017). Furthermore, even though the embodied cognition framework predicts interindividually different cognitive processing, comparatively little work has been done in that regard. Therefore, this finding is of high relevance both for the language production research community as well as for researchers investigating semantic processing in general.

Following from these findings, there are several promising research areas: First, interoception has been named an important candidate to explain the grounding of concepts which are less concrete (Connell et al., 2018; Lynott et al., 2020) and it is related to processing words referring to mental states, emotions or social concepts (Villani et al., 2021). Therefore, interindividual differences should be considered when investigating those concepts. Second, it will be interesting to compare different aspects of interoception and their role in semantic processing.

We chose two subscales from the MAIA questionnaire (Mehling et al., 2018, 2012) which in our eyes best measured sensitivity towards the experimental manipulation of body movements. In Experiment 2 from Study 2 we also used a mindfulness questionnaire (FFMQ: Baer et al., 2008). However, the traits which were assessed with those scales did not moderate the outcome of the sentence completion task. Potentially, more objective interoception measures like heartbeat tracking might be another promising avenue to take even though they do not measure interoceptive sensibility but interoceptive accuracy (Garfinkel et al., 2015). Finally, it should be investigated if differences in interoceptive abilities are linked to other aspects of communication and language production like "being good with words", talkative, etc.

Taken together, the results from Study 2 suggest that differences in a trait like interoceptive sensibility may moderate embodied lexical selection in an unconstrained language production task.

4.2.2 Musical expertise

Furthermore, there is tentative evidence that also different levels of expertise lead to differential semantic processing in language production. The evidence stems from a preliminary analysis of an unpublished dataset which has been collected within the scope of this dissertation project. In a PWI task, participants had to name objects which were either strongly perceived auditorily (e.g., <piano> or <rooster>) or were not related to strong auditory experiences (e.g., <snail> or <book>). Distractor words were neither categorically nor associatively related to the target words but were also either strongly perceived auditorily or were not related to auditory experiences. The Lancaster Sensorimotor Norms were used to control for the auditory experience of the stimulus set (Lynott et al., 2020) and we manipulated shared experiential traces between the target words and the distractor words. For example, the picture of a <trombone> (auditory related) was paired with the distractor word *fire enginge* (auditorily related) and in a different condition with the distractor word *cheese* (not auditorily related). The auditorily related target objects were either instruments (25 %), animals producing a typical sound (25 %) or other sound related objects (50 %). We wanted to know whether shared experiential traces between a target and a distractor word lead to similar semantic context effects as previously found in the PWI (Bürki et al., 2020). Both musicians and non-musicians (n = 24 in each group) took part in the study. We assumed that participants with musical expertise would be more sensitive towards auditory features in their environment: There is evidence that auditory experiences related to concepts are retrieved in language comprehension (e.g., Kiefer et al., 2008; Trumpp et al., 2014) and that music expertise may lead to differential processing of auditorily related content in language comprehension (Wolter et al., 2017). Therefore, an effect of shared auditory experiential traces between a target and a distractor might be restricted to musicians.

Preliminary analyses suggest that there was no general effect of a shared auditory experiential trace between the target pictures and the distractor words. However, there was a trend that musicians showed a semantic interference effect when they had to name instruments while ignoring sound-related distractor words which were not musical instruments. Importantly, it has previously been shown that visual presentation of music instruments leads to activation in auditory association areas in the brain for musicians but not for non-musicians (Hoenig et al., 2011). The preliminary experiential interference effect we found would thus speak for a reactivation of auditory experiential traces when music experts are naming instruments. In case sound related words had to be ignored, this led to interference in comparison to naming pictures of instruments with distractor words which are not associated with producing a sound. As the experiment was conducted online and only 24 musicians took part, power may not have been sufficient to draw sound statistical conclusions because previously 48 subjects were needed to replicate a semantic interference effect in an online PWI task (see Study 3). Therefore, more data will have to be collected in the future. If these preliminary results stabilize, this would show that interindividual differences due to expertise play out during semantic processing in language production. Besides it would show that experiential traces from the auditory domain are important for language production (see also Grechuta et al., 2020; Mädebach, Kieseler, & Jescheniak, 2018; Mulatti et al., 2014). On the other hand, shared experiential traces by themselves do not seem to elicit classic semantic context effects if participants are not sensitive towards this modality.

While Study 1 and 2 used a novel paradigm to investigate the content of lexical choices, carefully designed classic language production paradigms like the PWI will make it easier to link findings on embodied meaning aspects to the previous literature on semantic processing in language production. Preliminary evidence suggests that different levels of expertise may lead to differences in processing the experientially grounded meaning aspect of auditory strength as indicated by an interference effect in picture naming.

4.3 Integrating the findings in models of semantic processing

Theories of semantic processing in language production did not yet integrate experientially grounded theories or distributional accounts of meaning (Vinson et al., 2014). This makes it difficult to relate the above findings to the framework of semantic processing in language production laid out in Section 1.2.1 in the Introduction. However, it should be noted that the ignorance with regard to advances in both, empirical insights as well as theoretical reasoning is on both sides: language production theory did not integrate advances in theoretical reasoning about semantics which has mainly been informed by language comprehension research (see e.g., Bi, 2021; Binder & Desai, 2011; Davis & Yee, 2021; Kuhnke, Kiefer, & Hartwigsen, 2021; Lambon Ralph, 2014; Pulvermüller, 2018), and these frameworks did not integrate evidence from language production research for the most part. Below I will discuss some thoughts which have been sparked by the present findings and which may provide first steps towards a more comprehensive framework of semantic processing in language production.

4.3.1 Are experiential traces semantic features?

According to the framework of semantic processing in language production presented in the Introduction (Section 1.2.1), the activation of semantic features and the resulting interplay of co-activated lexical concepts and lemmas can account for semantic effects in language production. Traditionally, semantic features have been viewed as propositional knowledge which is abstracted from experience (see e.g., Günther et al., 2019). For example, the concept <mouse> may be described by the propositions IS(MOUSE, MAMMAL), HAS(MOUSE, TAIL), EATS(MOUSE, CHEESE). In other words, the traditional notion of semantic features sees them as amodal and abstract and at least historically, this is also how semantic features have been envisioned in language production (Bierwisch & Schreuder, 1992). This view is difficult to link to accounts of embodied or grounded cognition which view concepts as not being abstracted away from (sensory) experiences. For example, features such as GREY and SMALL are part of the semantic representation of the concept <mouse>, too. If they are described as amodal and abstract propositional knowledge, this entirely fails to account for what the nature of the feature GREY is (see also Meteyard et al., 2012).

Above, I proposed that shared experiential traces between an experimental manipulation and a produced word like *shell* lead to the observed effects. To integrate this explanation in a semantic architecture with propositional semantic features we would have to assume that the findings from Study 1 and 2 can be explained by an activation of abstract semantic features which denote the spatial characteristics of concepts' referents. Thus, the experiential trace of vertical space of a lexical concept like <shell> should be thought of as a semantic feature which is abstracted away from experiences like IS LOCATED(SHELL, DOWN). Consequently, the head movement would have activated an UP- or DOWN-feature as would the presented sentence and then both the movement and the spatial features of the sentences would have independently raised the chances that lexical concepts which share that feature will be selected as suitable sentence ending.

This kind of explanation was given as one possibility to account for the finding that naming verbs for depicted actions which are blocked according to their effector (hand-only actions vs. foot-only actions) is slowed in homogeneous compared to heterogeneous blocks. Hirschfeld and Zwitserlood (2012) proposed that naming action verbs activates abstract semantic features like EXECUTED WITH THE HANDS and that this might lead to semantic competition between words sharing this feature. However, this explanation would fail to explain that (experientially grounded) meaning can be flexible and dynamic. For example, the finding that the picture of a <car> results in different naming latencies depending on its movement speed (Ben-Haim et al., 2015) would be difficult to explain with an invariable semantic feature like FAST. Relatedly, the spatial properties of objects are not static and absolute: a plane can be on the runway or in the air, a head is vertically higher up than feet but certainly less than a flying helicopter. Therefore, explaining the findings from this dissertation with semantic features like UP or DOWN as they have traditionally been understood would be infelicitous in my regard.

Alternatively, and in line with theories of embodied language processing presented in Section 1.1.1 in the Introduction, semantic features can be envisioned as being flexibly activated, directly linked to sensory circuits in the brain and not as being abstract, static and amodal. Then, concepts are linked to different kinds of experiential traces according to the sum of our individual experiences. Depending on the conversational context, these traces might be more or less relevant. If features are understood in such a way, this should be made explicit because otherwise terminologies like *features* lose explanatory power and depending on each researcher's background may be interpreted differently.

While I do not think that abstract semantic features would be the most likely way to explain my findings as well as other findings on embodied language processing, it has to be admitted that there is an ongoing discussion to which degree semantic processing relies on modal and grounded vs. amodal and more abstract representations (e.g., Bi, 2021; Coccia, Bartolini, Luzzi, Provinciali, & Lambon Ralph, 2004). This is partly also due to little consensus about what modal and amodal representation formats entail (Michel, 2021).

Therefore, my findings highlight that an integration of experientially grounded meaning aspects in theories of semantic processing during language production should go along with theoretical work which carefully assesses the explanatory value of labels and terminology. Along these lines, care should be taken to account for the fact that some features may themselves form complex concepts whereas others might be more basic (for empirical work on this issue see e.g., Binder et al., 2016).

4.3.2 A new framework of semantic processing in language production?

To account for the findings from Study 1 and 2, a framework is needed which can explain which nouns are likely candidates to complete the sentence fragments given (1) the syntactic structure of the sentences, (2) the semantics of the previous words in the sentence and (3) the current bodily state of the participants. To achieve this, a prediction mechanism is needed which keeps track of the distributional properties of language because certain words will be more likely depending on previous words. In fact, it has been shown that the naming latencies of multi-word phrases decrease with higher frequency of the phrase irrespective of the frequency of the noun in the phrase (Janssen & Barber, 2012). This sensitivity of our language production system to the statistical distribution of words has not been integrated in models of semantic processing so far (Vinson et al., 2014). With regard to this dissertation, it is an important link because the distributional semantic similarity between the nouns in the presented sentences and the produced nouns was interacting with the spatial properties of the presented sentences.

To me, connectionist models seem a promising framework to explain and model the findings from this dissertation. They have been introduced into research on language production with regards to syntax and phonology (Dell, 1986, 1988) but semantics has been widely neglected (for exceptions see Calvillo, Brouwer, & Crocker, 2021; Oppenheim, 2018). To the best of my knowledge, connectionist language production models did not yet include those semantic aspects I have been concerned with in this dissertation.

They seem especially suitable for two reasons: (1) Computational implementations of connectionist models can incrementally predict upcoming words (Dell & Chang, 2014). This feature would be helpful to model the task in Study 1 and 2 where participants first had to understand the content of the sentence fragments in order to produce a suitable word. Thus, connectionist models provide a way to link comprehension and production, which are not only intertwined in the tasks used in the presented studies but in human communication in general (Pickering & Garrod, 2013). (2) Representations in connectionist models would be in line with experientially grounded views of concepts. According to McClelland and Cleeremans (2009), they are not thought of as propositional representations. They are construed of as an activation pattern over the units in a network, e.g. the representation of an object or a word is the sum of the activations while we perceive it. Each encounter with an object or word can leave a trace in this activation network and this in turn leads to an adjustment in the connections to other units. Via this learning mechanism, previous experiences can later be reconstructed and reactivated. This architecture is similar to the neural mechanisms suggested by embodied theories of semantic processing (e.g., Pulvermüller, 2018). Accordingly, semantic processing relies on activations of widely-distributed networks in the brain. It is especially the flexible and constant adjustment of the fiberways connecting neural areas via Hebbian learning which can explain conceptual processing.

As a next step, it would be interesting to see whether existing connectionist models of the lexical-semantic system which are based on a grounded cognition approach (e.g., by Ursino, Cuppini, & Magosso, 2010) can be adjusted to account for the findings in this dissertation as well as for other semantic context effects in language production. Such a model should also be able to incorporate that meanings can be constructed on the fly (see e.g., Lin et al., 2021) and that meaning processing may depend on interindividual experiences or traits like interoceptive sensibility (see Study 2).

Integrating the findings from this dissertation in a framework of semantic processing is a timely endevaour. In fact, there is an ongoing and lively debate on the architecture of our semantic memory and meaning processing mechanisms, which gave rise to the recent statement that "[w]e don't know how the brain stores anything, let alone words" (Poeppel & Idsardi, 2022, p. 1; but see e.g., Pulvermüller, 2018). The debate demonstrates the relevance of finding clever paradigms to investigate semantic processing (Niv, 2021; Ostarek & Bottini, 2020) as well as the need for rigorous thinking about the computational, representational, and implementational level of semantic processing in language comprehension and production (Borghi & Fini, 2019; Marr & Poggio, 1976; van Rooij & Baggio, 2021; Vinson et al., 2014).

Note that even though I aimed to provide conceptual clarity, the lack of such in some parts of this dissertation may be due to the vagueness which may result when interdisciplinary lines of research with different traditions of labeling and theorising are brought together as well as due to the subject matter and its unresolved issues.

4.3.3 The quest for alternative paradigms

The main scope of application of current frameworks for semantic processing in language production has been to account for and explain findings from picture naming studies with voice onset latencies as dependent variable (see Introduction, Section 1.2.1 and 1.2.2). It seems as if the field has focused on effects, while losing sight of the primary explanandum: the human capacity for speech, that is finding the right words to express an intended meaning. However, priority should be given not to explain effects like facilitation or interference which have been observed in experiments but to give explanations of real-world capacities (van Rooij & Baggio, 2021).

The use of picture naming tasks yielded many insights about human language processing and they became widely used because researcher did not want to be confined to study speech errors and were interested in the time course of language production (Levelt et al., 1999). Analysing differences in the timing of responses allowed inferences about underlying cognitive processes only when nuisance variables were controlled for. This was easier when every participant produced the same words. However, picture naming is certainly not the most relevant feature of language production in everyday life. For understanding the language production capacity in general, we therefore need additional paradigms which allow to focus on lexical selection while still being able to run quantitative analyses.

The experiments presented in Study 1 and 2 are a move in that direction: We were focusing on which words people choose in a rather unconstrained language production task. Thus, we targeted the human capacity to produce language by examining whether experientially grounded meaning is among the factors which influence the selection of words. It is unquestionable that the task which was used is still far away from the real-world capacity of free language production in a communicative setting. Still it contained important aspects of language production in every day life, which are missing in most picture naming studies: (1) an interplay of language comprehension and production (Meyer, Huettig, & Levelt, 2016; Pickering & Garrod, 2013), (2) production of words which are embedded in a meaningful sentential context, and (3) the free choice of one lexical alternative from among several potentially suitable candidates.

Studying sentence completions as I did in this dissertation may be one route to take. Similarly, it may be a viable approach to use verbal fluency tasks. For example, participants produce as many words beginning with a specific letter in a given time and subsequently their answers are rated according to several meaning aspects (Klostermann et al., 2022). Relatedly, participants can be asked to produce as many category members as possible in a predefined time. Recently, it was shown that the rank and the frequency of named category members can be independently predicted by the sensorimotor similarity of the category members to the category as well as by distributional measures of semantic similarity (Banks et al., 2021). Another option will be to use methods from improvisation theatre to elicit speech (Fjaellingsdal et al., 2020). All this should go along with investigating language production in communicative and social settings to make the research more ecologically valid (e.g., Brehm, Taschenberger, & Meyer, 2019; Gambi et al., 2015; Kuhlen & Abdel Rahman, 2022; Lin et al., 2021).

4.4 How to move language production studies online

Due to the pandemic, researchers were forced to think of alternatives to lab-based in-person testing. Up until then, language production research relying on overt auditory responses had been confined to the lab. This was due to problems with audiovisual synchrony which resulted in poor reliability of presenting visual and auditory stimuli at the same time in web-based experiments (Bridges et al., 2020; Reimers & Stewart, 2016). To the best of our knowledge none of the available softwares for running online experiments solved the audiovisual synchrony problem so far. However, with Study 3, a proof of concept for running online language production experiments using voice onset latencies from picture naming tasks as dependent measure is provided. To this end, we replicated a PWI experiment which had previously been conducted in the lab and compared naming latencies and response times for vocal-consonant classifications of the last letter of a target word (Abdel Rahman & Aristei, 2010). We showed that semantic interference effects can reliably be obtained in both tasks and that they are comparable in size to lab-based experiments (Bürki et al., 2020).

Based on our experiences several suggestions can be made for running online (language production) experiments, which will raise the data quality while minimizing the resources for testing on the side of the researchers. First, the web experiment builder should be carefully chosen. Data from Experiment 2 and 3 suggest that using the open source software jspsych (de Leeuw, 2015) yielded more reliable results than data from Experiment 1 collected using the experiment platform SoSciSurvey (Leiner, 2019). Second, carefully testing the experiment in different configurations using different browsers and devices will help to minimize noise stemming from the technical side. Third, noise stemming from the participant side should be reduced by carefully thinking about the recruitment procedure, by giving ample feedback to participants and by including catch trials. Furthermore, researchers should account for potentially increased levels of noise in their experiment by raising the sample size or the number of trials. Moreover, the data should be carefully screened before analysis and criteria for the exclusion of trials and participants should be preregistered (see also Rodd, 2021).

In the future, the range of research questions which can be investigated will hopefully be getting wider. This can already be seen in Study 2 from this dissertation which used the online audio recording implementation to record responses in a design where careful monitoring of participants was needed. To this end, a video connection with the experimenter was established while running the experiment. Thereby, we successfully showed that online language production studies can be run even if a high level of control is needed.

Subsequent work from our lab and from other researchers also confirmed the viability of running online language studies with overt auditory responses (Fairs & Strijkers, 2021; Li et al., 2022; Stark et al., 2022). This raises the chances that language production research will not be confined to laboratories any more. By moving research to the internet, larger and more diverse samples may be reached, too. Testing hitherto understudied populations like speakers of minority languages (Speed et al., 2018) or clinical samples with reduced mobility (Stark, 2022) will be a great chance to validate empirical findings in language production.

4.5 Conclusions and future directions

We still do not understand how humans solve the ubiquitious yet cognitively complicated task of selecting the right words to express an extended meaning. The present studies are an important step for gaining a better understanding of this cognitive capacity. I investigated the lexical content of language production by moving beyond picture naming tasks with Study 1 and 2, which has not been done widely so far. The findings support the view that lexical choices can be influenced by the reactivation of experiential traces. Thereby, the sparse literature on experientially grounded meaning aspects in language production is extended.

As also shown in these two studies, experiential reactivation can come in different flavours: directly grounded meanings can be targeted when asking participants to engage in a body movement and additionally, language itself can serve as a means to reactivate experiences. When we read or hear about certain situations this also reactivates experiential traces which can contribute independently to our lexical choices. Thus, the findings highlight the importance of investigating more meaning dimensions in language production. Furthermore, interindividual differences in interoceptive sensibility interacted with embodied semantic processing. Therefore, the role of interindividual differences for language production should receive more attention in the future.

Even though the reported effects have been small, they can lead to the selection of one word instead of another one due to the reactivation of experiential traces. This may have huge implications for real life conversations because the use of different lexical items due to our bodily state can result in a completely different communicative outcome.

Interpreting these empirical insights with reference to theories of semantic memory, it seems evident that concepts in language production are flexible and based on experiences which can come from both direct interactions of our bodies with the world as well as from exposure to language. However, there is no encompassing theoretical framework of semantic processing in language production where these findings fit in. To move forward, it will be necessary to carefully assess the terminology of the available literature. Rigorous theoretical work is needed which should be complemented by more diverse, cleverly designed and contextually rich experimental paradigms.

References

Abdel Rahman, R., & Aristei, S. (2010). Now you see it . . . and now again: Semantic interference reflects lexical competition in speech production with and without articulation. *Psychonomic Bulletin & Review*, *17*(5), 657–661. doi: 10.3758/PBR.17.5.657

Abdel Rahman, R., & Melinger, A. (2007). When bees hamper the production of honey: Lexical interference from associates in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 604–614. doi: 10.1037/0278-7393.33.3.604

Abdel Rahman, R., & Melinger, A. (2009). Semantic context effects in language production: A swinging lexical network proposal and a review. *Language and Cognitive Processes*, *24*(5), 713–734. doi: 10.1080/01690960802597250

Abdel Rahman, R., & Melinger, A. (2011). The dynamic microstructure of speech production: Semantic interference built on the fly. *Journal of Experimental Psychology: Learning Memory and Cognition*, *37*(1), 149–161. doi: 10.1037/a0021208

Abdel Rahman, R., & Melinger, A. (2019). Semantic processing during language production: An update of the swinging lexical network. *Language, Cognition and Neuroscience*, *34*(9), 1176–1192. doi: 10.1080/23273798.2019.1599970

Alario, F. X., Segui, J., & Ferrand, L. (2000). Semantic and associative priming in picture naming. *The Quarterly Journal of Experimental Psychology Section A*, *53*(3), 741–764. doi: 10.1080/713755907

Andrews, M., Frank, S., & Vigliocco, G. (2014). Reconciling embodied and distributional accounts of meaning in language. *Topics in Cognitive Science*, *6*(3), 359–370. doi: 10.1111/tops.12096

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, *53*, 1407–1425. doi: 10.3758/s13428-020-01501-5

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388–407. doi: 10.3758/s13428-019-01237-x

Aravena, P., Courson, M., Frak, V., Cheylus, A., Paulignan, Y., Deprez, V., & Nazir, T. A. (2014). Action relevance in linguistic context drives word-induced motor activity. *Frontiers in Human Neuroscience*, *8*, Article 163. doi: 10.3389/fnhum.2014.00163

Aristei, S., & Abdel Rahman, R. (2013). Semantic interference in language production is due to graded similarity, not response relevance. *Acta Psychologica*, *144*(3), 571–582. doi: 10.1016/j.actpsy.2013.09.006

Aristei, S., Melinger, A., & Abdel Rahman, R. (2011). Electrophysiological chronometry of semantic context effects in language production. *Journal of Cognitive Neuroscience*, *23*(7), 1567–1586. doi: 10.1162/jocn.2010.21474

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28. doi: 10.1109/TMAG.1979.1060222

Baer, R. A., Smith, G. T., Lykins, E., Button, D., Krietemeyer, J., Sauer, S., … Williams, J. M. G. (2008). Construct validity of the five facet mindfulness questionnaire in meditating and nonmeditating samples. *Assessment*, *15*(3), 329–342. doi: 10.1177/1073191107313003

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459. doi: 10.3758/BF03193014

Banks, B., Wingfield, C., & Connell, L. (2021). Linguistic distributional knowledge and sensorimotor grounding both contribute to semantic category production. *Cognitive Science*, *45*(10), Article e13055. doi: 10.1111/cogs.13055

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The waCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226. doi: 10.1007/s10579-009-9081-4

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, *16*(7), 419–429. doi: 10.1038/nrn3950

Barrós-Loscertales, A., González, J., Pulvermüller, F., Ventura-Campos, N., Bustamante, J. C., Costumero, V., . . . Ávila, C. (2012). Reading salt activates gustatory brain regions: fMRI evidence for semantic grounding in a novel sensory modality. *Cerebral Cortex*, *22*(11), 2554–2563. doi: 10.1093/cercor/bhr324

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*(3), 211–227. doi: 10.3758/BF03196968

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660. doi: 10.1017/S0140525X99002149

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645. doi: 10.1146/annurev.psych.59.103006.093639

Barsalou, L. W. (2020). Challenges and opportunities for grounding cognition. *Journal of Cognition*, *3*(1), 1–24. doi: 10.5334/joc.116

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *ArXiV*. doi: https://doi.org/10.48550/arXiv.1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Beilock, S. L., Lyons, I. M., Mattarella-Micke, A., Nusbaum, H. C., & Small, S. L. (2008). Sports experience changes the neural processing of action language. *Proceedings of the*

National Academy of Sciences of the United States of America, *105*(36), 13269–13273. doi: 10.1073/pnas.0803424105

Belke, E. (2013). Long-lasting inhibitory semantic context effects on object naming are necessarily conceptually mediated: Implications for models of lexical-semantic encoding. *Journal of Memory and Language*, *69*(3), 228–256. doi: 10.1016/j.jml.2013.05.008

Belke, E., Meyer, A. S., & Damian, M. F. (2005). Refractory effects in picture naming as assessed in a semantic blocking paradigm. *The Quarterly Journal of Experimental Psychology Section A*, *58*(4), 667–692. doi: 10.1080/02724980443000142

Belke, E., & Stielow, A. (2013). Cumulative and non-cumulative semantic interference in object naming: Evidence from blocked and continuous manipulations of semantic context. *Quarterly Journal of Experimental Psychology*, *66*(11), 2135–2160. doi: 10.1080/17470218.2013.775318

Bellmund, J. L., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, *362*(6415). doi: 10.1126/science.aat6766

Ben-Haim, M. S., Chajut, E., Hassin, R. R., & Algom, D. (2015). Speeded naming or naming speed? The automatic effect of object speed on performance. *Journal of Experimental Psychology: General*, *144*(2), 326–338. doi: 10.1037/a0038569

Bergen, B. K. (2015). Embodiment. In E. Dabrowska & D. Divjak (Eds.), *Handbook of cognitive linguistics* (pp. 10–30). Berlin: De Gruyter.

Bergen, B. K., Lindsay, S., Matlock, T., & Narayanan, S. (2007). Spatial and linguistic aspects of visual imagery in sentence comprehension. *Cognitive Science*, *31*(5), 733–764. doi: 10.1080/03640210701530748

Berndt, E., Dudschig, C., & Kaup, B. (2018). Activating concepts by activating experiential traces: Investigations with a series of anagram solution tasks. *The Quarterly Journal of Experimental Psychology*, *71*(2), 483–498. doi: 10.1080/17470218.2016.1261913

Bi, Y. (2021). Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences*, *25*(10), 883–895. doi: 10.1016/j.tics.2021.07.006

Bierwisch, M., & Schreuder, R. (1992). From concepts to lexical items. *Cognition*, *42*(1-3), 23–60. doi: 10.1016/0010-0277(92)90039-K

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, *33*(3-4), 130–174. doi: 10.1080/02643294.2016.1147426

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. doi: 10.1016/j.tics.2011.10.001.

Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research Methods*, *42*(3), 665–670. doi: 10.3758/BRM.42.3.665

Boersma, P., & Weenink, D. (2020). Praat: Doing phonetics by computer [Computer software manual]. Retrieved from `http://www.praat.org` (Version: 6.1.37)

Bonin, P., & Fayol, M. (2000). Writing words from pictures: What representations are activated, and when? *Memory and Cognition*, *28*(4), 677–689. doi: 10.3758/BF03201257

Borghi, A. M., & Fini, C. (2019). Theories and explanations in psychology. *Frontiers in Psychology*, *10*(APR), 1–3. doi: 10.3389/fpsyg.2019.00958

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243. doi: 10.1111/j.2517-6161.1964.tb00553.x

Brand, A., & Bradley, M. T. (2012). Assessing the effects of technical variance on the statistical outcomes of web experiments measuring response times. *Social Science Computer Review*, *30*(3), 350–357. doi: 10.1177/0894439311415604

Brehm, L., Taschenberger, L., & Meyer, A. (2019). Mental representations of partner task cause interference in picture naming. *Acta Psychologica*, *199*, 102888. doi: 10.1016/J.ACTPSY.2019.102888

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*. doi: 10.7717/peerj.9414

Buldeo, N. (2015). Interoception: A measure of embodiment or attention? *International Body Psychotherapy Journal*, *14*(1), 65–79.

Bürki, A., Elbuy, S., Madec, S., & Vasishth, S. (2020). What did we learn from forty years of research on semantic interference? A Bayesian meta-analysis. *Journal of Memory and Language*, *114*, Article 104125. doi: 10.1016/j.jml.2020.104125

Calvillo, J., Brouwer, H., & Crocker, M. W. (2021). Semantic systematicity in connectionist language production. *Information*, *12*(8), 329. doi: 10.3390/info12080329

Calvo-Merino, B., Grèzes, J., Glaser, D. E., Passingham, R. E., & Haggard, P. (2006). Seeing or doing? Influence of visual and motor familiarity in action observation. *Current Biology*, *16*(19), 1905–1910. doi: 10.1016/j.cub.2006.07.065

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, *14*(1), 177–208. doi: 10.1080/026432997381664

Carota, F., Nili, H., Pulvermüller, F., & Kriegeskorte, N. (2021). Distinct fronto-temporal substrates of distributional and taxonomic similarity among words: Evidence from RSA of BOLD signals. *NeuroImage*, *224*, Article 117408. doi: 10.1016/j.neuroimage.2020.117408

Carota, F., Schoffelen, J. M., Oostenveld, R., & Indefrey, P. (2022). The time course of language production as revealed by pattern classification of MEG sensor data. *Journal of Neuroscience*, *42*(29), 5745–5754. doi: 10.1523/JNEUROSCI.1923-21.2022

Casasanto, D., & Henetz, T. (2012). Handedness shapes children's abstract concepts. *Cognitive Science*, *36*(2), 359–372. doi: 10.1111/j.1551-6709.2011.01199.x

Cicchetti, D. V., Shoinralter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement*, *9*(1), 31–36. doi: 10.1177/014662168500900103

Coccia, M., Bartolini, M., Luzzi, S., Provinciali, L., & Lambon Ralph, M. A. (2004). Semantic memory is an amodal, dynamic system: Evidence from the interaction of naming and object use in semantic dementia. *Cognitive Neuropsychology*, *21*(5), 513–527. doi: 10.1080/02643290342000113

Connell, L. (2007). Representing object colour in language comprehension. *Cognition*, *102*(3), 476–485. doi: 10.1016/j.cognition.2006.02.009

Connell, L., Lynott, D., & Banks, B. (2018). Interoception: The forgotten modality in perceptual grounding of abstract and concrete concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752). doi: 10.1098/rstb.2017.0143

Costa, A., Alario, F.-X., & Caramazza, A. (2005). On the categorical nature of the semantic interference effect in the picture-word interference paradigm. *Psychonomic Bulletin & Review*, *12*(1), 125–131. doi: 10.3758/BF03196357

Costa, A., Strijkers, K., Martin, C., & Thierry, G. (2009). The time course of word retrieval revealed by event-related brain potentials during overt speech. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(50), 21442–21446. doi: 10.1073/pnas.0908921106

Cotelli, M., Borroni, B., Manenti, R., Zanetti, M., Arévalo, A., Cappa, S. F., & Padovani, A. (2007). Action and object naming in Parkinson's disease without dementia. *European Journal of Neurology*, *14*(6), 632–637. doi: 10.1111/j.1468-1331.2007.01797.x

Craig, A. D. (2002). How do you feel? Interoception: The sense of the physiological condition of the body. *Nature Reviews Neuroscience*, *3*(8), 655–666. doi: 10.1038/nrn894

Damian, M. F., & Spalek, K. (2014). Processing different kinds of semantic relations in picture-word interference with non-masked and masked distractors. *Frontiers in Psychology*, *5*, Article 1183. doi: 10.3389/fpsyg.2014.01183

Damian, M. F., Vigliocco, G., & Levelt, W. J. (2001). Effects of semantic context in the naming of pictures and words. *Cognition*, *81*(3), 77–86. doi: 10.1016/S0010-0277(01)00135-4

Davis, C. P., & Yee, E. (2021). Building semantic memory from embodied and distributional language experience. *WIREs Cognitive Science*, *12*(5), Article e1555. doi: 10.31234/OSF.IO/WYMR9

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12. doi: 10.3758/s13428-014-0458-y

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321. doi: 10.1037/0033-295X.93.3.283

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, *27*(2), 124–142. doi: 10.1016/0749-596X(88)90070-8

Dell, G. S., & Chang, F. (2014). The p-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634). doi: 10.1098/rstb.2012.0394

de Zubicaray, G., Fraser, D., Ramajoo, K., & McMahon, K. (2017). Interference from related actions in spoken word production: Behavioural and fMRI evidence. *Neuropsychologia*, *96*, 78–88. doi: 10.1016/j.neuropsychologia.2017.01.010

de Zubicaray, G., Hansen, S., & McMahon, K. (2013). Differential processing of thematic and categorical conceptual relations in spoken word production. *Journal of Experimental Psychology: General*, *142*(1), 131–142. doi: 10.1037/a0028717

de Zubicaray, G., McLean, M., Oppermann, F., Hegarty, A., McMahon, K., & Jescheniak, J. D. (2018). The shape of things to come in speech production: Visual form interference during lexical access. *Quarterly Journal of Experimental Psychology*, *71*(9), 1921–1938. doi: 10.1080/17470218.2017.1367018

Diamond, M. (2016). Recorderjs [Computer software manual]. Retrieved from `https://github.com/mattdiamond/Recorderjs`

Díez-Álamo, A. M., Diéz, E., Wojcik, D. Z., Alonso, M. A., & Fernandez, A. (2018). Sensory experience ratings for 5,500 Spanish words. *Behavior Research Methods*, *51*(3), 1205–1215. doi: 10.3758/s13428-018-1057-0

Dijkstra, K., Kaschak, M. P., & Zwaan, R. A. (2007). Body posture facilitates retrieval of autobiographical memories. *Cognition*, *102*(1), 139–149. doi: 10.1016/j.cognition.2005.12.009

Dreyer, F. R., Frey, D., Arana, S., von Saldern, S., Picht, T., Vajkoczy, P., & Pulvermüller, F. (2015). Is the motor system necessary for processing action and abstract emotion words? Evidence from focal brain lesions. *Frontiers in Psychology*, *6*, 1–17. doi: 10.3389/fpsyg.2015.01661

Dudschig, C., & Kaup, B. (2017). Is it all task-specific? The role of binary responses, verbal mediation, and saliency for eliciting language-space associations. *Journal of Experimental Psychology: Learning Memory and Cognition*, *43*(2), 259–270. doi: 10.1037/xlm0000297

Dudschig, C., Lachmair, M., de la Vega, I., De Filippis, M., & Kaup, B. (2012). From top to bottom: Spatial shifts of attention caused by linguistic stimuli. *Cognitive Processing*, *13*(Suppl. 1), 151–154. doi: 10.1007/s10339-012-0480-x

Dudschig, C., Souman, J., & Kaup, B. (2013). Motion in vision and language: Seeing visual motion can influence processing of motion verbs. In *Proceedings of the 35th annual conference of the Cognitive Science Society* (pp. 2225–2230). Cognitive Science Society.

Durda, K., Buchanan, L., & Caron, R. (2009). Grounding co-occurrence: Identifying features in a lexical co-occurrence model of semantic memory. *Behavior Research Methods*, *41*(4), 1210–1223. doi: 10.3758/BRM.41.4.1210

Estes, Z., Verges, M., & Barsalou, L. W. (2008). Head up, foot down: Object words orient attention to the objects' typical location. *Psychological Science*, *19*(2), 93–97. doi: 10.1111/j.1467-9280.2008.02051.x.

Fairs, A., & Strijkers, K. (2021). Can we use the internet to study speech production? Yes we can! Evidence contrasting online versus laboratory naming latencies and errors. *PLoS*

ONE, *16*(10), e0258908. doi: 10.1371/journal.pone.0258908

Fargier, R., Montant, M., & Strijkers, K. (2019). The activation of sensory and emotional experience during speech production. *Poster presented at NeuroFrance - Annual meeting of the French Society for Neuroscience.*

Finn, R. (1972). Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement*, *32*(2), 255–265. doi: 10.1177/001316447203200203

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-55. In *Studies in linguistic analysis* (pp. 1–32). Oxford: Blackwell.

Fjaellingsdal, T. G., Schwenke, D., Scherbaum, S., Kuhlen, A. K., Bögels, S., Meekes, J., & Bleichner, M. G. (2020). Expectancy effects in the EEG during joint and spontaneous word-by-word sentence production in German. *Scientific Reports*, *10*, Article 5460. doi: 10.1101/782581

Fleiss, J. L. (1986). *The design and analysis of clinical experiments.* New York: Wiley. doi: 10.1002/bimj.4710300308

Fodor, J. A. (1975). *The language of thought.* Harvard: Harvard University Press.

Gallant, J., & Libben, G. (2019). No lab, no problem. Designing lexical comprehension and production experiments using PsychoPy3. *The Mental Lexicon*, *14*(1), 152–168. doi: 10.1075/ml.00002.gal

Gambi, C., Van de Cavey, J., & Pickering, M. J. (2015). Interference in joint picture naming. *Journal of Experimental Psychology: Learning Memory and Cognition*, *41*(1), 1–21. doi: 10.1037/a0037438

García, A. M., Moguilner, S., Torquati, K., García-Marco, E., Herrera, E., Muñoz, E., … Ibáñez, A. (2019). How meaning unfolds in neural time: Embodied reactivations can precede multimodal semantic effects during language processing. *Neuroimage*, *197*, 439–449. doi: 10.1016/j.neuroimage.2019.05.002

Garfinkel, S. N., Seth, A. K., Barrett, A. B., Suzuki, K., & Critchley, H. D. (2015). Knowing your own heart: Distinguishing interoceptive accuracy from interoceptive awareness. *Biological Psychology*, *104*, 65–74. doi: 10.1016/j.biopsycho.2014.11.004

Gatti, D., Marelli, M., Vecchi, T., & Rinaldi, L. (2022). Spatial representations without spatial computations. *Psychological Science*, Advance online. doi: 10.1177/09567976221094863

Gibson, J. (2019). Mindfulness, interoception, and the body: A contemporary perspective. *Frontiers in Psychology*, *10*, 2012. doi: 10.3389/fpsyg.2019.02012

Gilbert, B. (2020). jspsych-image-audio-response.js [Computer software manual]. Retrieved from `https://github.com/becky-gilbert/jsPsych/blob/audio-response/docs/plugins/jspsych-image-audio-response.md`

Gilbert, B., & Minors, D. (2020). audio-response-timing [Computer software manual]. Retrieved from `https://github.com/becky-gilbert/audio-response-timing`

Glaser, W. R., & Düngelhoff, F.-J. (1984). The time course of picture-word interference. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(5), 640–654. doi: 10.1037//0096-1523.10.5.640.

Glenberg, A. M. (2010). Embodiment as a unifying perspective for psychology. *WIREs Cognitive Science*, *1*(4), 586–596. doi: 10.1002/wcs.55

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, *9*(3), 558–565. doi: 10.1002/bit

González, J., Barros-Loscertales, A., Pulvermüller, F., Meseguer, V., Sanjuán, A., Belloch, V., & Ávila, C. (2006). Reading cinnamon activates olfactory brain regions. *Neuroimage*, *32*(2), 906–912. doi: 10.1016/j.neuroimage.2006.03.037

Gozli, D. G., Chasteen, A. L., & Pratt, J. (2013). The cost and benefit of implicit spatial cues for visual attention. *Journal of Experimental Psychology: General*, *142*(4), 1028–1046. doi: 10.1037/a0030362

Grechuta, K., Rubio Ballester, B., Espín Munné, R., Usabiaga Bernal, T., Molina Hervás, B., Mohr, B., … Verschure, P. F. (2020). Multisensory cueing facilitates naming in aphasia. *Journal of NeuroEngineering and Rehabilitation*, *17*, Article 122. doi: 10.1186/s12984-020-00751-w

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. doi: 10.1111/2041-210X.12504

Grootswagers, T. (2020). A primer on running human behavioural experiments online. *Behavior Research Methods*, *52*(6), 2283–2286. doi: 10.3758/s13428-020-01395-3

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun - An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*, *47*(4), 930–944. doi: 10.3758/s13428-014-0529-0

Günther, F., Dudschig, C., & Kaup, B. (2018). Symbol grounding without direct experience: Do words inherit sensorimotor activation from purely linguistic context? *Cognitive Science*, *42*(Suppl 2), 336–374. doi: 10.1111/cogs.12549

Günther, F., Nguyen, T., Chen, L., Dudschig, C., Kaup, B., & Glenberg, A. M. (2020). Immediate sensorimotor grounding of novel concepts learned from language alone. *Journal of Memory and Language*, *115*, Article 104172. doi: 10.1016/j.jml.2020.104172

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033. doi: 10.1177/1745691619861372

Häfner, M. (2013). When body and mind are talking: Interoception moderates embodied cognition. *Experimental Psychology*, *60*(4), 255–259. doi: 10.1027/1618-3169/a000194

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. doi: 10.20982/tqmp.08.1.p023.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335–346. doi: 10.1016/0167-2789(90)90087-6

Hartmann, M., Grabherr, L., & Mast, F. W. (2012). Moving along the mental number line: Interactions between whole-body motion and numerical cognition. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(6), 1416–1427. doi: 10.1037/a0026706

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*(2), 301–307. doi: 10.1016/S0896-6273(03)00838-9

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. doi: 10.1017/S0140525X0999152X

Herbert, B. M., & Pollatos, O. (2012). The body in the mind: On the relationship between interoception and embodiment. *Topics in Cognitive Science*, *4*(4), 692–704. doi: 10.1111/j.1756-8765.2012.01189.x

Herrera, E., Rodríguez-Ferreiro, J., & Cuetos, F. (2012). The effect of motion content in action naming by Parkinson's disease patients. *Cortex*, *48*(7), 900–904. doi: 10.1016/j.cortex.2010.12.007

Hilbig, B. E. (2016). Reaction time effects in lab- versus web-based research: Experimental evidence. *Behavior Research Methods*, *48*(4), 1718–1724. doi: 10.3758/s13428-015-0678-9

Hirschfeld, G., & Zwitserlood, P. (2012). Effector-specific motor activation modulates verb production. *Neuroscience Letters*, *523*(1), 15–18. doi: 10.1016/j.neulet.2012.06.025

Hoenig, K., Müller, C., Herrnberger, B., Sim, E.-J., Spitzer, M., Ehret, G., & Kiefer, M. (2011). Neuroplasticity of semantic representations for musical instruments in professional musicians. *NeuroImage*, *56*(3), 1714–1725. doi: 10.1016/J.NEUROIMAGE.2011.02.065

Hoenig, K., Sim, E.-J., Bochev, V., Herrnberger, B., & Kiefer, M. (2008). Conceptual flexibility in the human brain: Dynamic recruitment of semantic maps from visual, motor, and motion-related areas. *Journal of Cognitive Neuroscience*, *20*(10), 1799–1814. doi: 10.1162/

jocn.2008.20123

Holt, L. E., & Beilock, S. L. (2006). Expertise and its embodiment: Examining the impact of sensorimotor skill expertise on the representation of action-related text. *Psychonomic Bulletin & Review*, *13*(4), 694–701. doi: 10.3758/BF03193983

Howard, D., Nickels, L., Coltheart, M., & Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: Experimental and computational studies. *Cognition*, *100*(3), 464–482. doi: 10.1016/j.cognition.2005.02.006

Hughes, J. W., & Schnur, T. T. (2017). Facilitation and interference in naming: A consequence of the same learning process? *Cognition*, *165*, 61–72. doi: 10.1016/j.cognition.2017.04.012

Hustá, C., Zheng, X., Papoutsi, C., & Piai, V. (2021). Electrophysiological signatures of conceptual and lexical retrieval from semantic memory. *Neuropsychologia*, *161*, 107988. doi: 10.1016/j.neuropsychologia.2021.107988

Hutson, J., Damian, M. F., & Spalek, K. (2013). Distractor frequency effects in picture-word interference tasks with vocal and manual responses. *Language and Cognitive Processes*, *28*(5), 615–632. doi: 10.1080/01690965.2011.605599

Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology*, *2*, 1–16. doi: 10.3389/fpsyg.2011.00255

Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1-2), 101–144. doi: 10.1016/j.cognition.2002.06.001

Jackson, R. L., Hoffman, P., Pobric, G., & Lambon Ralph, M. A. (2015). The nature and neural correlates of semantic association versus conceptual similarity. *Cerebral Cortex*, *25*(11), 4319–4333. doi: 10.1093/cercor/bhv003

Janssen, N., & Barber, H. A. (2012). Phrase frequency effects in language production. *PLoS ONE*, *7*(3), Article e33202. doi: 10.1371/journal.pone.0033202

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, *20*(4), 824–843. doi: 10.1037/0278-7393.20.4.824

Kabat-Zinn, J. (1990). *Full catastrophe living: The program of the stress reduction clinic at the University of Massachusetts Medical Center.*

Kaschak, M. P., Madden, C. J., Therriault, D. J., Yaxley, R., Aveyard, M., Blanchard, A. A., & Zwaan, R. A. (2005). Perception of motion affects language processing. *Cognition*, *94*(3), B79–B80. doi: 10.1016/j.cognition.2004.06.005

Kaup, B., de la Vega, I., Strozyk, J., & Dudschig, C. (2016). The role of sensorimotor processes in meaning composition. In M. H. Fischer & Y. Coello (Eds.), *Conceptual and interactive embodiment: Foundations of embodied cognition* (Vol. 2, pp. 46–66). London: Routledge. doi: 10.4324/9781315751962

Keehner, M., & Fischer, M. H. (2012). Unusual bodies, uncommon behaviors: Individual and group differences in embodied cognition in spatial tasks. *Spatial Cognition and Computation*, *12*(2-3), 71–82. doi: 10.1080/13875868.2012.659303

Kelter, S., & Kaup, B. (2012). Conceptual knowledge, categorization, and meaning. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An international handbook of natural language meaning* (pp. 2775–2804). Berlin: De Gruyter. doi: 10.1515/9783110589825-011

Kemmerer, D. (2006). The semantics of space: Integrating linguistic typology and cognitive neuroscience. *Neuropsychologia*, *44*(9), 1607–1621. doi: 10.1016/j.neuropsychologia.2006.01.025

Khan, M. (2020). RecordRTC [Computer software manual]. Retrieved from `https://recordrtc.org`

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex*, *48*(7), 805–825. doi: 10.1016/j.cortex.2011.04.006

Kiefer, M., Sim, E.-J., Herrnberger, B., Grothe, J., & Hoenig, K. (2008). The sound of concepts: Four markers for a link between auditory and conceptual brain systems. *The Journal of Neuroscience*, *28*(47), 12224–12230. doi: 10.1523/JNEUROSCI.3579-08.2008

Klostermann, F., Wyrobnik, M., Boll, M., Ehlen, F., & Tiedt, H. O. (2022). Tracing embodied word production in persons with Parkinson's disease in distinct motor conditions. *Scientific Reports*, *12*(1), Article 16669. doi: 10.1038/s41598-022-21106-6

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. doi: 10.1016/j.jcm.2016.02.012

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, *33*(2), 149–174. doi: 10.1006/jmla.1994.1008

Kuhlen, A. K., & Abdel Rahman, R. (2017). Having a task partner affects lexical retrieval: Spoken word production in shared task settings. *Cognition*, *166*, 94–106. doi: 10.1016/j.cognition.2017.05.024

Kuhlen, A. K., & Abdel Rahman, R. (2022). Mental chronometry of speaking in dialogue: Semantic interference turns into facilitation. *Cognition*, *219*, Article 104962. doi: 10.1016/j.cognition.2021.104962

Kuhnke, P., Kiefer, M., & Hartwigsen, G. (2021). Task-dependent functional and effective connectivity during conceptual processing. *Cerebral Cortex*, *31*, 3475–3493. doi: 10.1093/cercor/bhab026

Kumle, L., Võ, M. L., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, *53*, 2528–2543. doi: 10.3758/s13428-021-01546-0

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi: 10.18637/

JSS.V082.I13

La Heij, W., Dirkx, J., & Kramer, P. (1990). Categorical interference and associative priming in picture naming. *British Journal of Psychology*, *81*(4), 511–525. doi: 10.1111/j.2044-8295.1990.tb02376.x

Lachmair, M., Dudschig, C., De Filippis, M., de la Vega, I., & Kaup, B. (2011). Root versus roof: Automatic activation of location information during word processing. *Psychonomic Bulletin & Review*, *18*(6), 1180–1188. doi: 10.3758/s13423-011-0158-x

Lachmair, M., Dudschig, C., de la Vega, I., & Kaup, B. (2016). Constructing meaning for up and down situated sentences: Is a sentence more than the sum of its words? *Language and Cognition*, *8*(4), 604–628. doi: 10.1017/langcog.2015.11

Lachmair, M., Ruiz Fernández, S., Bury, N.-A., Gerjets, P., Fischer, M. H., & Bock, O. L. (2016). How body orientation affects concepts of space, time and valence: Functional relevance of integrating sensorimotor experiences during word processing. *PLoS ONE*, *11*(11), Article e0165795. doi: 10.1371/journal.pone.0165795

Lambon Ralph, M. A. (2014). Neurocognitive insights on conceptual knowledge and its breakdown. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634). doi: 10.1098/rstb.2012.0392

Lange, K., Kühn, S., & Filevich, E. (2015). "Just another tool for online studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLoS ONE*, *10*(6), Article e0130834. doi: 10.1371/journal.pone.0130834

Leiner, D. J. (2019). SoSci Survey [Computer software manual]. Retrieved from https://www.soscisurvey.de/de/about (Version 3.1.06)

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge: Cambridge University Press.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75. doi: https://doi.org/10.1017/

S0140525X99001776

Li, K. K., Schwarz, J., Sim, J. H., Zhang, Y., Buchanan-Worster, E., Post, B., & McDougall, K. (2022). Recording and timing vocal responses in online experimentation. In *Interspeech proceedings 2022* (pp. 4053–4057). International Speech Communication Association. doi: 10.21437/Interspeech.2022-10697

Liepelt, R., Dolk, T., & Prinz, W. (2012). Bidirectional semantic interference between action and speech. *Psychological Research*, *76*(4), 446–455. doi: 10.1007/s00426-011-0390-z

Lin, H.-P., Kuhlen, A. K., & Abdel Rahman, R. (2021). Ad-hoc thematic relations form through communication: Effects on lexical-semantic processing during language production. *Language, Cognition and Neuroscience*, *36*(9), 1057–1075. doi: 10.1080/23273798.2021.1900580

Lin, H.-P., Kuhlen, A. K., & Abdel Rahman, R. (2022). Robust cumulative semantic interference for (very) loose semantic relations in the continuous naming paradigm. *Manuscript submitted for publication*, Humboldt–Universität zu Berlin.

Lorenz, A., Regel, S., Zwitserlood, P., & Abdel Rahman, R. (2018). Age-related effects in compound production: Intact lexical representations but more effortful encoding. *Acta Psychologica*, *191*, 289–309. doi: 10.1016/j.actpsy.2018.09.001

Louwerse, M. M. (2008). Embodied relations are encoded in language. *Psychonomic Bulletin & Review*, *15*(4), 838–844. doi: 10.3758/PBR.15.4.838

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273–302. doi: 10.1111/j.1756-8765.2010.01106.x

Louwerse, M. M. (2018). Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Topics in Cognitive Science*, *10*(3), 573–589. doi: 10.1111/tops.12349

Louwerse, M. M., Hutchinson, S., Tillman, R., & Recchia, G. (2015). Effect size matters: The role of language statistics and perceptual simulation in conceptual processing. *Language, Cognition and Neuroscience*, *30*(4), 430–447. doi: 10.1080/23273798.2014.981552

Louwerse, M. M., & Zwaan, R. A. (2009). Language encodes geographical information. *Cognitive Science*, *33*(1), 51–73. doi: 10.1111/j.1551-6709.2008.01003.x

Lupker, S. J. (1979). The semantic nature of response competition in the picture-word interference task. *Memory & Cognition*, *7*(6), 485–495. doi: 10.3758/BF03198265

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster sensorimotor norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, *52*(3), 1271–1291. doi: 10.3758/s13428-019-01316-z

Mädebach, A., Kieseler, M. L., & Jescheniak, J. D. (2018). Localizing semantic interference from distractor sounds in picture naming: A dual-task study. *Psychonomic Bulletin & Review*, *25*(5), 1909–1916. doi: 10.3758/s13423-017-1386-5

Mädebach, A., Wöhner, S., Kieseler, M.-L., & Jescheniak, J. D. (2017). Neighing, barking, and drumming horses — Object related sounds help and hinder picture naming. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(9), 1629–1646. doi: 10.1037/xhp0000415

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology Paris*, *102*(1-3), 59–70. doi: 10.1016/j.jphysparis.2008.03.004

Mahon, B. Z., Costa, A., Peterson, R., Vargas, K. A., & Caramazza, A. (2007). Lexical selection is not by competition: A reinterpretation of semantic interference and facilitation effects in the picture-word interference paradigm. *Journal of Experimental Psychology: Learning Memory and Cognition*, *33*(3), 503–535. doi: 10.1037/0278-7393.33.3.503

Marr, D., & Poggio, T. (1976). From understanding computation to understanding neural circuitry. *AI Memos*, *357*.

Matheson, H. E., White, N., & McMullen, P. A. (2014). Testing the embodied account of object naming: A concurrent motor task affects naming artifacts and animals. *Acta Psychologica*, *145*(1), 33–43. doi: 10.1016/j.actpsy.2013.10.012

Mathôt, S., Grainger, J., & Strijkers, K. (2017). Pupillary responses to words that convey a sense of brightness or darkness. *Psychological Science*, *28*(8), 1116–1124. doi: 10.1177/0956797617702699

Mathot, S., & March, J. (2022). Conducting linguistic experiments online with OpenSesame and OSWeb. *Language Learning*, Advance online. doi: https://doi.org/10.1111/lang.12509

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. doi: 10.1016/j.jml.2017.01.001

McClelland, J. L., & Cleeremans, A. (2009). Connectionist models. In T. Byrne, A. Cleeremans, & P. Wilken (Eds.), *Oxford companion to consciousness.* Oxford: Oxford University Press.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*, 547–559. doi: 10.3758/bf03192726

Medin, D. L., Ross, N. O., Atran, S., Cox, D., Coley, J., Proffitt, J. B., & Blok, S. (2006). Folkbiology of freshwater fish. *Cognition*, *99*(3), 237–273. doi: 10.1016/j.cognition.2003.12.005

Mehling, W. E., Acree, M., Stewart, A., Silas, J., & Jones, A. (2018). The multidimensional assessment of interoceptive awareness, Version 2 (MAIA-2). *PLoS ONE*, *13*(12), e0208034. doi: 10.1371/journal.pone.0208034

Mehling, W. E., Price, C., Daubenmier, J. J., Acree, M., Bartmess, E., & Stewart, A. (2012). The multidimensional assessment of interoceptive awareness (MAIA). *PLoS ONE*, *7*(11), e48230. doi: 10.1371/journal.pone.0048230

Meteyard, L., Cuadrado Rodriguez, S., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, *48*, 788–804. doi: 10.1016/j.cortex.2010.11.002

Meteyard, L., Zokaei, N., Bahrami, B., & Vigliocco, G. (2008). Visual motion interferes with lexical decision on motion words. *Current Biology*, *18*(17), 732–733. doi: 10.1016/j.cub.2008.07.016

Meyer, A. S., Huettig, F., & Levelt, W. J. (2016). Same, different, or closely related: What is the relationship between language production and comprehension? *Journal of Memory and Language*, *89*, 1–7. doi: 10.1016/j.jml.2016.03.002

Michalak, J., Zarbock, G., Drews, M., Otto, D., Mertens, D., Ströhle, G., … Heidenreich, T. (2016). Erfassung von Achtsamkeit mit der deutschen Version des Five Facet Mindfulness Questionnaires (FFMQ-D). *Zeitschrift für Gesundheitspsychologie*, *24*(1), 1–12. doi: 10.1026/0943-8149/a000149

Michel, C. (2021). Overcoming the modal/amodal dichotomy of concepts. *Phenomenology and the Cognitive Sciences*, *20*(4), 655–677. doi: 10.1007/s11097-020-09678-y

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiV*. doi: 10.48550/arXiv.1301.3781

Morey, R. D., Kaschak, M. P., Díez-Álamo, A. M., Glenberg, A. M., Zwaan, R. A., Lakens, D., … Ziv-Crispel, N. (2022). A pre-registered, multi-lab non-replication of the action-sentence compatibility effect (ACE). *Psychonomic Bulletin & Review*, *29*(2), 613–626. doi: 10.3758/s13423-021-01927-8

Mulatti, C., Treccani, B., & Job, R. (2014). The role of the sound of objects in object identification: Evidence from picture naming. *Frontiers in Psychology*, *5*, Article 1139. doi: 10.3389/fpsyg.2014.01139

Niv, Y. (2021). The primacy of behavioral research for understanding the brain. *Behavioral Neuroscience*, *135*(5), 601–609. doi: 10.1037/bne0000471

Oaster, T. R. F. (1989). Number of alternatives per choice point and stability of Likert-type scales. *Perceptual and Motor Skills*, *68*(2), 549–550. doi: 10.2466/pms.1989.68.2.549

Oppenheim, G. M. (2018). The paca that roared: Immediate cumulative semantic interference among newly acquired words. *Cognition*, *177*, 21–29. doi: 10.1016/j.cognition.2018.02.014

Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, *114*(2), 227–252. doi: 10.1016/j.cognition.2009.09.007

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867–872. doi: 10.1016/j.jesp.2009.03.009

Ostarek, M. (2018). *Envisioning language — An exploration of perceptual processes in language comprehension* (Dissertation). Radboud Universiteit Nijmegen.

Ostarek, M., & Bottini, R. (2020). Towards strong inference in research on embodiment — Possibilities and limitations of causal paradigms. *Journal of Cognition*, *4*(1), 1–21. doi: 10.5334/joc.139

Ostarek, M., & Huettig, F. (2019). Six challenges for embodiment research. *Current Directions in Psychological Science*, *28*(6), 593–599. doi: 10.1177/0963721419866441

Ostarek, M., Ishag, A., Joosen, D., & Huettig, F. (2018). Saccade trajectories reveal dynamic interactions of semantic and spatial information during the processing of implicitly spatial words. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *44*(10), 1658–1670. doi: 10.1037/xlm0000536

Ostarek, M., Joosen, D., Ishag, A., de Nijs, M., & Huettig, F. (2019). Are visual processes causally involved in "perceptual simulation" effects in the sentence-picture verification task? *Cognition*, *182*, 84–94. doi: 10.1016/j.cognition.2018.08.017

Ostarek, M., Van Paridon, J., & Huettig, F. (2018). Cross-decoding reveals shared brain activity patterns between saccadic eye-movements and semantic processing of implicitly spatial words. *bioRxiv*, 415596. doi: 10.1101/415596

Ostarek, M., & Vigliocco, G. (2017). Reading sky and seeing a cloud: On the relevance of events for perceptual simulation. *Journal of Experimental Psychology: Learning Memory* *and Cognition*, *43*(4), 579–590. doi: 10.1037/xlm0000318

Öttl, B., Dudschig, C., & Kaup, B. (2017). Forming associations between language and sensorimotor traces during novel word learning. *Language and Cognition*, *9*(1), 156–171. doi: 10.1017/langcog.2016.5

Papesh, M. H. (2015). Just out of reach: On the reliability of the action-sentence compatibility effect. *Journal of Experimental Psychology: General*, *145*(6), e116–e141. doi: 10.1037/xge0000218

Pecher, D., van Dantzig, S., & Schifferstein, H. N. J. (2009). Concepts are not represented by conscious imagery. *Psychonomic Bulletin & Review*, *16*(5), 914–919. doi: 10.3758/PBR.16.5.914

Pecher, D., & Zwaan, R. A. (2005). *Grounding cognition. The role of perception and action in memory, language, and thinking* (Vol. 26) (No. 11). Cambridge: Cambridge University Press. doi: 10.1016/j.patrec.2005.01.006

Piai, V., Roelofs, A., & Van Der Meij, R. (2012). Event-related potentials and oscillatory brain responses associated with semantic and Stroop-like interference effects in overt naming. *Brain Research*, *1450*, 87–101. doi: 10.1016/j.brainres.2012.02.050

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347. doi: 10.1017/S0140525X12001495

Pinet, S., Zielinski, C., Mathôt, S., Dufau, S., Alario, F. X., & Longcamp, M. (2017). Measuring sequences of keystrokes with jsPsych: Reliability of response times and interkeystroke intervals. *Behavior Research Methods*, *49*(3), 1163–1176. doi: 10.3758/s13428-016-0776-3

Pitt, B., & Casasanto, D. (2019). The correlations in experience principle: How culture shapes concepts of time and number. *Journal of Experimental Psychology: General*, *149*(6). doi: 10.1037/xge0000696

Plant, R. R. (2016). A reminder on millisecond timing accuracy and potential replication failure in computer-based psychology experiments: An open letter. *Behavior Research Methods*, *48*(1), 408–411. doi: 10.3758/s13428-015-0577-0

Pobric, G., Jefferies, E., & Lambon Ralph, M. A. (2010a). Category-specific versus category-general semantic impairment induced by transcranial magnetic stimulation. *Current Biology*, *20*(10), 964–968. doi: 10.1016/J.CUB.2010.03.070

Pobric, G., Jefferies, E., & Lambon Ralph, M. A. (2010b). Induction of semantic impairments using rTMS: Evidence for the hub-and-spoke semantic theory. *Behavioural Neurology*, *23*(4), 217–219. doi: 10.3233/BEN-2010-0299

Poeppel, D., & Idsardi, W. (2022). We don't know how the brain stores anything, let alone words. *Trends in Cognitive Sciences*. doi: 10.1016/j.tics.2022.08.010

Poldrack, R. A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., ... Bilder, R. M. (2011). The cognitive atlas: Toward a knowledge foundation for cognitive neuroscience. *Frontiers in Neuroinformatics*, *5*, 17. doi: 10.3389/fninf.2011.00017

Pulvermüller, F. (2013). Semantic embodiment, disembodiment or misembodiment? In search of meaning in modules and neuron circuits. *Brain and Language*, *127*(1), 86–103. doi: 10.1016/j.bandl.2013.05.015

Pulvermüller, F. (2018). Neural reuse of action perception circuits for language, concepts and communication. *Progress in Neurobiology*, *160*, 1–44. doi: 10.1016/j.pneurobio.2017.07.001

Pulvermüller, F., Hauk, O., Nikulin, V. V., & Ilmoniemi, R. J. (2005). Functional links between motor and language systems. *European Journal of Neuroscience*, *21*(3), 793–797. doi: 10.1111/j.1460-9568.2005.03900.x

Pulvermüller, F., Shtyrov, Y., & Hauk, O. (2009). Understanding in an instant: Neurophysiological evidence for mechanistic language circuits in the brain. *Brain and Language*, *110*(2), 81–94. doi: 10.1016/J.BANDL.2008.12.001

Ramsay, J. O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika*, *38*(4), 513–532. doi: 10.1007/BF02291492

Redmann, A., FitzPatrick, I., Hellwig, F., & Indefrey, P. (2014). The use of conceptual components in language production: An ERP study. *Frontiers in Psychology*, *5*, Article 363. doi: 10.3389/fpsyg.2014.00363

Reifegerste, J., Meyer, A. S., Zwitserlood, P., & Ullman, M. T. (2021). Aging affects steaks more than knives: Evidence that the processing of words related to motor skills is relatively spared in aging. *Brain and Language*, *218*. doi: 10.1016/j.bandl.2021.104941

Reimers, S., & Stewart, N. (2016). Auditory presentation and synchronization in Adobe Flash and HTML5/JavaScript web experiments. *Behavior Research Methods*, *48*(3), 897–908. doi: 10.3758/s13428-016-0758-5

Revelle, W. (2018). psych: procedures for psychological, psychometric, and personality research [Computer software manual]. (R package version 1.8.12)

Richardson, D. C., Spivey, M. J., Barsalou, L. W., & McRae, K. (2003). Spatial representations activated during real-time comprehension of verbs. *Cognitive Science*, *27*(5), 767–780. doi: 10.1016/S0364-0213(03)00064-8

Rodd, J. (2021). Collecting experimental data online: How to maintain data quality when you can't see your participants. *The Journal of the Acoustical Society of America*, *149*(4), A81–A81. doi: 10.1121/10.0004581

Rodríguez-Ferreiro, J., Menéndez, M., Ribacoba, R., & Cuetos, F. (2009). Action naming is impaired in Parkinson disease patients. *Neuropsychologia*, *47*(14), 3271–3274. doi: 10.1016/j.neuropsychologia.2009.07.007

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, *42*(1-3), 107–142. doi: 10.1016/0010-0277(92)90041-F

Roelofs, A. (2018). A unified computational account of cumulative semantic, semantic blocking, and semantic distractor effects in picture naming. *Cognition*, *172*, 59–72. doi: 10.1016/j.cognition.2017.12.007

Roelofs, A., Piai, V., & Schriefers, H. (2013). Context effects and selective attention in picture naming and word reading: Competition versus response exclusion. *Language and Cognitive Processes*, *28*(5), 655–671. doi: 10.1080/01690965.2011.615663

Rose, S. B., & Abdel Rahman, R. (2016). Cumulative semantic interference for associative relations in language production. *Cognition*, *152*, 20–31. doi: 10.1016/j.cognition.2016.03.013

Rose, S. B., & Abdel Rahman, R. (2017). Semantic similarity promotes interference in the continuous naming paradigm: Behavioural and electrophysiological evidence. *Language, Cognition and Neuroscience*, *32*(1), 55–68. doi: 10.1080/23273798.2016.1212081

Rose, S. B., Aristei, S., Melinger, A., & Abdel Rahman, R. (2019). The closer they are, the more they interfere: Semantic similarity of word distractors increases competition in language production. *Journal of Experimental Psychology: Learning Memory and Cognition*, *45*(4), 753–763. doi: 10.1037/xlm0000592

Roux, F., Armstrong, B. C., & Carreiras, M. (2017). Chronset: An automated tool for detecting speech onset. *Behavior Research Methods*, *49*(5), 1864–1881. doi: 10.3758/s13428-016-0830-1

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, *20*(1), 33–53.

Samuelson, L. K., Smith, L. B., Perry, L. K., & Spencer, J. P. (2011). Grounding word learning in space. *PLoS ONE*, *6*(12). doi: 10.1371/journal.pone.0028095

Sauter, M., Draschkow, D., & Mack, W. (2020). Building, hosting and recruiting: A brief introduction to running behavioral experiments online. *Brain Sciences*, *10*(4), 1–11. doi: 10.3390/BRAINSCI10040251

Schriefers, H., Meyer, A. S., & Levelt, W. J. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language*, *29*(1), 86–102. doi: 10.1016/0749-596X(90)90011-N

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow norms:

Ratings of 5,500 words on nine scales. *Behavior Research Methods*, *51*(3), 1258–1270. doi: 10.3758/s13428-018-1099-3

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–457. doi: 10.1017/S0140525X00005756

Seger, B. T., Hauf, J. E., & Nieding, G. (2020). Perceptual simulation of vertical object movement during comprehension of auditory and audiovisual text in children and adults. *Discourse Processes*, *57*(5-6), 460–472. doi: 10.1080/0163853X.2020.1755801

Shao, Z., van Paridon, J., Poletiek, F., & Meyer, A. S. (2019). Effects of phrase and word frequencies in noun phrase production. *Journal of Experimental Psychology: Learning Memory and Cognition*, *45*(1), 147–165. doi: 10.1037/xlm0000570

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. doi: 10.1037/0033-2909.86.2.420

Sixtus, E., Lindemann, O., & Fischer, M. H. (2018). Incidental counting: Speeded number naming through finger movements. *Journal of Cognition*, *1*(1), Article 44. doi: 10.5334/joc.49

Speed, L. J., & Majid, A. (2019). Grounding language in the neglected senses of touch, taste, and smell. *Cognitive Neuropsychology*, *37*(5-6), 363–392. doi: 10.1080/02643294.2019.1623188

Speed, L. J., Vinson, D. P., & Vigliocco, G. (2015). Representing meaning. In E. Dabrowska & D. Divjak (Eds.), *Handbook of cognitive linguistics* (pp. 190–211). Berlin: De Gruyter.

Speed, L. J., Wnuk, E., & Majid, A. (2018). Studying psycholinguistics out of the lab. In A. M. B. de Groot & P. Hagoort (Eds.), *Research methods in psycholinguistics and the neurobiology of language: A practical guide* (pp. 190–207). New York: Wiley.

Stark, K. (2022). *Perspectives for online research: Examples from speech production and social settings.* Presentation at Université de Lille: ESCoP Conference.

Stark, K., van Scherpenberg, C., Obrig, H., & Abdel Rahman, R. (2022). Web-based language production experiments: Semantic interference assessment is robust for spoken and typed response modalities. *Behavior Research Methods*, Advance online. doi: 10.3758/s13428-021-01768-2

Strijkers, K., & Costa, A. (2016). The cortical dynamics of speaking: Present shortcomings and future avenues. *Language, Cognition and Neuroscience*, *31*(4), 484–503. doi: 10.1080/23273798.2015.1120878

Strozyk, J. V., Dudschig, C., & Kaup, B. (2019). Do I need to have my hands free to understand hand-related language? Investigating the functional relevance of experiential simulations. *Psychological Research*, *83*(3), 406–418. doi: 10.1007/s00426-017-0900-8

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., … Bates, E. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language*, *51*(2), 247–250. doi: 10.1016/j.jml.2004.03.002

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, *23*(3), 457–482. doi: 10.1016/0010-0285(91)90016-H

Thornton, T., Loetscher, T., Yates, M. J., & Nicholls, M. E. (2013). The highs and lows of the interaction between word meaning and space. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(4), 964–973. doi: 10.1037/a0030467

Torrance, M., Nottbusch, G., Alves, R. A., Arfé, B., Chanquoy, L., Chukharev-Hudilainen, E., … Wengelin, Å. (2018). Timed written picture naming in 14 European languages. *Behavior Research Methods*, *50*(2), 744–758. doi: 10.3758/s13428-017-0902-x

Trevethan, R. (2017). Intraclass correlation coefficients: Clearing the air, extending some cautions, and making some requests. *Health Services and Outcomes Research Methodology*, *17*(2), 127–143. doi: 10.1007/s10742-016-0156-6

Trumpp, N. M., Kliese, D., Hoenig, K., Haarmeier, T., & Kiefer, M. (2013). Losing the sound of concepts: Damage to auditory association cortex impairs the processing of sound-related concepts. *Cortex*, *49*(2), 474–486. doi: 10.1016/j.cortex.2012.02.002

Trumpp, N. M., Traub, F., Pulvermüller, F., & Kiefer, M. (2014). Unconscious automatic brain activation of acoustic and action-related conceptual features during masked repetition priming. *Journal of Cognitive Neuroscience*, *26*(2), 352–364. doi: 10.1162/jocn_a_00473

Tufft, M. R. A., & Richardson, D. C. (2020). Social offloading: Just working together is enough to remove semantic interference. In *Proceedings of the 42th annual meeting of the Cognitive Science Society* (pp. 859–865). Cognitive Science Society.

Ulrich, R., & Giray, M. (1989). Time resolution of clocks: Effects on reaction time measurement — Good news for bad clocks. *British Journal of Mathematical and Statistical Psychology*, *42*(1), 1–12. doi: 10.1111/j.2044-8317.1989.tb01111.x

Unger, L., & Fisher, A. V. (2021). The emergence of richly organized semantic knowledge from simple statistics: A synthetic review. *Developmental Review*, *60*, Article 100949. doi: 10.1016/j.dr.2021.100949

Ursino, M., Cuppini, C., & Magosso, E. (2010). A computational model of the lexical-semantic system based on a grounded cognition approach. *Frontiers in Psychology*, *1*. doi: 10.3389/fpsyg.2010.00221

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*(4), 682–697. doi: 10.1177/1745691620970604

Verges, M., & Duffy, S. (2009). Spatial representations elicit dual-coding effects in mental imagery. *Cognitive Science*, *33*(6), 1157–1172. doi: 10.1111/j.1551-6709.2009.01038.x

Vieth, H. E., McMahon, K. L., & de Zubicaray, G. I. (2014). Feature overlap slows lexical selection: Evidence from the picture-word interference paradigm. *Quarterly Journal of Experimental Psychology*, *67*(12), 2325–2339. doi: 10.1080/17470218.2014.923922

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition*, *1*(2), 219–247. doi: 10.1515/langcog.2009.011

Vigliocco, G., Vinson, D. P., Damian, M. F., & Levelt, W. (2002). Semantic distance effects on object and action naming. *Cognition*, *85*(3), B61–B69. doi: 10.1016/S0010-0277(02)00107-5

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, *48*(4), 422–488. doi: 10.1016/j.cogpsych.2003.09.001

Villani, C., Lugli, L., Liuzza, M. T., Nicoletti, R., & Borghi, A. M. (2021). Sensorimotor and interoceptive dimensions in concrete and abstract concepts. *Journal of Memory and Language*, *116*, Article 104173. doi: 10.1016/j.jml.2020.104173

Vincent-Lamarre, P., Massé, A. B., Lopes, M., Lord, M., Marcotte, O., & Harnad, S. (2015). The latent structure of dictionaries. *Topics in Cognitive Science*, *8*(3), 625–659.

Vinson, D. P., Andrews, M., & Vigliocco, G. (2014). Giving words meaning: Why better models of semantics are needed in language production research. In M. Goldrick, V. S. F. Ferreira, & M. Miozzo (Eds.), *The Oxford handbook of language production* (pp. 134–151). Oxford: Oxford University Press.

Vinson, D. P., & Vigliocco, G. (2002). A semantic analysis of grammatical class impairments: Semantic representations of object nouns, action nouns and action verbs. *Journal of Neurolinguistics*, *15*(3-5), 317–351. doi: 10.1016/S0911-6044(01)00037-9

Voeten, C. C. (2021). buildmer: Stepwise elimination and term reordering for mixed-effects regression [Computer software manual]. Retrieved from `https://cran.r-project.org/web/packages/buildmer/buildmer.pdf` (Version 1.9)

Vogt, A., Ganter, I., Kaup, B., & Abdel Rahman, R. (2022). Embodied language production: Sensorimotor activations and interoceptive sensibility influence which words we choose when speaking. *PsyArXiv*. doi: 10.31234/osf.io/3zgrc

Vogt, A., Hauber, R., Kuhlen, A. K., & Abdel Rahman, R. (2022). Internet-based language production research with overt articulation: Proof of concept, challenges, and practical advice. *Behavior Research Methods*, *54*, 1954–1975. doi: 10.3758/s13428-021-01686-3

Vogt, A., Kaup, B., & Abdel Rahman, R. (2022). Experience-driven meaning affects lexical choices during language production. *Quarterly Journal of Experimental Psychology*, Advance online. doi: https://doi.org/10.1177/17470218221125425

Vogt, A., Kaup, B., & Dudschig, C. (2019). When words are upside down: Language–space associations in children and adults. *Journal of Experimental Child Psychology*, *186*, 142–158. doi: 10.1016/j.jecp.2019.06.001

Wellsby, M., & Pexman, P. M. (2014). Developing embodied cognition: Insights from children's concepts and language processing. *Frontiers in Psychology*, *5*, 1–10. doi: 10.3389/fpsyg.2014.00506

Willems, R. M., Labruna, L., D'Esposito, M., Ivry, R., & Casasanto, D. (2011). A functional role for the motor system in language understanding: Evidence from theta-burst transcranial magnetic stimulation. *Psychological Science*, *22*(7), 849–854. doi: 10.1177/0956797611412387

Wingfield, C., & Connell, L. (2022). Understanding the role of linguistic distributional knowledge in cognition. *Language, Cognition and Neuroscience*, Advance online. doi: https://doi.org/10.1080/23273798.2022.2069278

Winter, A., Dudschig, C., Miller, J., Ulrich, R., & Kaup, B. (2022). The action-sentence compatibility effect (ACE): Meta-analysis of a benchmark finding for embodiment. *Acta Psychologica*, *230*, Article 103712. doi: 10.1016/j.actpsy.2022.103712

Winter, B., & Matlock, T. (2013). More is up... and right: Random number generation along two axes. In *Proceedings of the 35th annual conference of the Cognitive Science Society* (pp. 3789–3974). Cognitive Science Society.

Witt, J. K., Kemmerer, D., Linkenauger, S. A., & Culham, J. (2010). A functional role for motor simulation in identifying tools. *Psychological Science*, *21*(9), 1215–1219. doi: 10.1177/0956797610378307

Wolter, S., Dudschig, C., & Kaup, B. (2017). Reading sentences describing high- or low-pitched auditory events: Only pianists show evidence for a horizontal space-pitch association. *Psychological Research*, *81*(6), 1213–1223. doi: 10.1007/s00426-016-0812-z

Yee, E., Chrysikou, E. G., Hoffman, E., & Thompson-Schill, S. L. (2013). Manual experience shapes object representations. *Psychological Science*, *24*(6), 909–919. doi: 10.1177/0956797612464658

Yee, E., & Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychonomic Bulletin & Review*, *23*(4), 1015–1027. doi: 10.3758/s13423-015-0948-7

Zwaan, R. A., & Madden, C. J. (2005). Embodied sentence comprehension. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thinking* (pp. 224–245). Cambridge: Cambridge University Press. doi: 10.1002/acp.1193

Zwaan, R. A., & Pecher, D. (2012). Revisiting mental simulation in language comprehension: six replication attempts. *PLoS ONE*, *7*(12), e51382. doi: 10.1371/journal.pone.0051382

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185. doi: 10.1037/0033-2909.123.2.162

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science*, *13*(2), 168–171. doi: 10.1111/1467-9280.00430

Zwaan, R. A., & Yaxley, R. H. (2003). Hemispheric differences in semantic-relatedness judgments. *Cognition*, *87*(3), B79–B86. doi: 10.1016/S0010-0277(02)00235-4

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt,

1. dass keine Zusammenarbeit mit gewerblichen Promotionsberatern stattfand,
2. dass ich die Dissertation auf der Grundlage der angegebenen Hilfsmittel und Hilfen selbstständig angefertigt habe,
3. dass ich mich nicht anderwärts um einen Doktorgrad beworben habe bzw. einen entsprechenden Doktorgrad besitze,
4. dass mir die dem angestrebten Verfahren zugrunde liegende Promotionsordnung der Lebenswissenschaftlichen Fakultät vom 05. März 2015, veröffentlicht im Amtlichen Mitteilungsblatt der Humboldt-Universität zu Berlin Nr. 12/2015 bekannt ist,
5. dass die Dissertation oder Teile davon nicht bereits bei einer anderen wissenschaftlichen Einrichtung eingereicht, angenommen oder abgelehnt wurden,
6. dass die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis eingehalten wurden.

Berlin, den