Symbol grounding without direct experience: Do words inherit sensorimotor activation from purely linguistic context?

Fritz Günther, Carolin Dudschig and Barbara Kaup
Department of Psychology, University of Tübingen, Germany

Author Note

This project was supported by the Collaborative Research Centre 833 (SFB 833) "The Construction of Meaning"/Z2 project appointed to Sigrid Beck and Barbara Kaup by the German Research Foundation (DFG). We thank Martin Lachmair and Birgit Öttl for providing us with their experimental data, which made possible important analyses in the preparation of our studies. We also thank our student assistants Florian Holl, Fabian Hurler, Dina Lausch, Eva Link, Janine Mayer and Lena Naumann for their valuable help in collecting the data.
Abstract

Theories of embodied cognition assume that concepts are grounded in non-linguistic, sensorimotor experience. In support of this assumption, previous studies have shown that upwards response movements are faster than downwards movements after participants have been presented with words whose referents are typically located in the upper vertical space (and vice versa for downwards responses). This is taken as evidence that processing these words re-activate sensorimotor experiential traces. This congruency effect was also found for novel words, after participants learned these words as labels for novel objects that they encountered either in their upper or lower visual field. While this indicates that direct experience with a word’s referent is sufficient to evoke said congruency effects, the present study investigates whether this direct experience is also a necessary condition. To this end, we conducted five experiments in which participants learned novel words from purely linguistic input: Novel words were presented in pairs with real up- or down-words (Experiment 1); they were presented in natural sentences where they replaced these real words (Experiment 2); they were presented as new labels for these real words (Experiment 3), and they were presented as labels for novel combined concepts based on these real words (Experiment 4 and 5). In all five experiments, we did not find any congruency effects elicited by the novel words; however, participants were always able to make correct explicit judgements about the vertical dimension associated to the novel words. These results suggest that direct experience is necessary for re-activating experiential traces, but this re-activation is not a necessary condition for understanding (in the sense of storing and accessing) the corresponding aspects of word meaning.

Keywords: Embodied Cognition; Language Comprehension; Language-Space Associations; Word Learning; Distributional Learning
Symbol grounding without direct experience: Do words inherit sensorimotor activation from purely linguistic context?

A remarkable and powerful cognitive skill that distinguishes humans from the vast majority of other animals is the capability to use symbols, and to engage in symbolic thought (Harnad, 1990; Hummel, 2010). Symbols are usually defined as arbitrary tokens that are semantically interpretable and can be manipulated and combined on the basis of rules. The most prominent symbol system available to us is natural language; other examples are artificial languages such as mathematics, programming languages and formal logic. In natural languages, words serve as symbols. In the human mind, the corresponding concepts are thought to be represented in a symbolic format as “mental symbols”, or symbolic representations, with which symbolic manipulations can be performed (Hummel, 2010; Murphy, 2002a).

For the last two decades, there has been an extensive debate over the nature and format of symbolic representations in the human mind (see Masson, 2015; Glenberg, 2015a, 2015b; Mahon, 2015a, 2015b, for an example and summary of this debate). Many classical studies have - sometimes implicitly - assumed that concepts are represented in our minds in an abstract, arbitrary and amodal way (e.g. J. R. Anderson, 1983; Collins & Quillian, 1969; Kintsch & van Dijk, 1978; Kintsch, 1988). The word “horse” has no obvious similarity to any horses in the world (highlighted by the fact that the same entities are referred to as “Pferd” in German and “cavallo” in Italian), and the same is thought to be the case for the concept, or mental symbol, HORSE - after all, the concept is a mental entity, and not an entity in the outside world. Therefore, concepts are often thought to be represented in our minds in an abstract format (Mahon, 2015a).

There are different classes of theories in which concepts are seen as such abstract, amodal symbols. Examples for such theories are semantic network models (Collins & Quillian, 1969; Kintsch, 1988; Quillian, 1967), distributional models of semantics (Landauer & Dumais, 1997; Lund & Burgess, 1996), or feature-based approaches to
semantics (McRae, Cree, Seidenberg, & McNorgan, 2005; Smith & Medin, 1981). For example, in a feature-based approach, the concept HORSE can be described by propositions such as \text{IS}(\text{HORSE, MAMMAL}), \text{HAS}(\text{HORSE, MANE}) \text{ and } \text{EATS}(\text{HORSE, HAY}). Critically, according to all of these theories, concepts are always defined by other concepts (in our example MAMMAL, MANE and HAY), to which they have some kind of relation or similarity, or which are their features.

However, such an approach to concepts does not come without problems. One argument that is often brought up against the assumption that mental symbols are represented in an amodal and abstract manner is the symbol grounding problem, often illustrated with the Chinese-Chinese dictionary argument (Harnad, 1990), which is derived from the Chinese room argument (Searle, 1980). In this thought experiment, a learner has the task of learning Chinese, with the only tool available to her being a monolingual Chinese dictionary. If she wants to look up the meaning of one specific symbol, it is defined via other symbols unknown to her, which in turn are defined via other unknown symbols - an infinite regress, with the result that no sensible meaning can ever be assigned to any symbol.

From this argument, it follows that symbols must be grounded in some way to obtain meaning, that is to be semantically interpretable. A prominent solution to the symbol grounding problem in cognitive psychology is the embodiment view, also known by the notion of grounded cognition. This view puts forward the argument that symbols ultimately have to be grounded in sensorimotor experience (Barsalou, 1999; Glenberg & Kaschak, 2002). The basic assumption of this account is that higher-order cognitive systems, such as language and the conceptual system, are not independent of perception and action and are therefore not abstract and amodal; instead, they use the same

1These terms are often used synonymously, but they can also be differentiated: Following the terminology proposed by Fischer (2012), effects of groundedness are based on universal facts and constraints of the physical world in general, while embodiedness effects are based on constraints put forward by individual sensory and motor learning experience and therefore can be affected training.
representation format and rely on the same cognitive systems. According to the embodiment view, language comprehension is a process of simulating the content of the language input with the same systems that are used for perception, action, and emotion (Glenberg, 2015a; Johnson-Laird, 1983). Strictly speaking, according to this view, the concept \textit{horse} does still not resemble real horses, as the concept \textit{horse} still is a mental entity and horses are entities in the outside world; however, the concept is similar to the \textit{experience} we make with real horses (or, more specifically, it is represented in the same format in the same systems). Hearing and thinking of horses, according to this view, activates the same systems as seeing horses, hearing them neigh, and the bodily sensations of riding a horse.

How are such embodied representations of concepts acquired? To address this question, Zwaan and Madden (2005) proposed the \textit{experiential trace model} of embodied language acquisition and comprehension. This model postulates that the driving factor for this acquisition is \textit{co-occurrence} of perceptual input and actions on the one hand, and linguistic input on the other hand (see also Hebb, 1949). When we see a galloping horse, and hear the word “horse” during this process, an associations is established between the visual experiential trace of the horse and the experiential trace for the linguistic construction “horse”. Upon hearing the word “horse” again in another situation, it serves as a cue to (partially) re-activate this visual experiential trace (for a computational model based on these assumptions of the experiential trace model, see Johns & Jones, 2015). This re-activation of perceptual input and actions is what enables us to understand the respective concept, by grounding it in sensorimotor experience.

Empirical evidence for the experiential trace model was provided in a study by Lachmair, Dudschig, De Filippis, de la Vega, and Kaup (2011). These authors showed that, upon encountering a word whose referent is typically associated with a specific vertical location in the world such as “roof” (an up-word) or “worm” (a down-word), experiential traces including this location feature are automatically re-activated. This re-activation
facilitated compatible vertical hand movements when responding towards the words in a variety of tasks, such as a Lexical Decision Task, or a Stroop Task where participants had to react according to the colour a word was printed in. Converging findings replicating these results were obtained by Thornton, Loetscher, Yates, and Nicholls (2013) (but see Dudschig & Kaup, 2017; Lebois, Wilson-Mendenhall, & Barsalou, 2015, for a more critical perspectives).

More direct evidence that such findings can indeed be attributed to experiential traces of perception and action that were encoded together with linguistic input comes from a recent study by Öttl, Dudschig, and Kaup (2017). In this study, participants had to learn novel word labels for novel objects. These objects were presented to them in the form of cuddly toys, either in their upper or lower visual field. After the correct labels had been learned, participants performed a test phase in the fashion of the Stroop Task by Lachmair et al. (2011), in which they responded according to the colour the novel words were printed in with either an upward or a downward movement. As in the study by Lachmair et al. (2011), a congruency effect was observed, with facilitated reactions if the learned vertical position of the novel word referents matched the response direction.

It is however important to assess the implications of the study by Öttl et al. (2017) for the experiential trace model. Since both completely novel words and novel objects were presented, with which no prior experience could have been made, the experience made in the learning phase must have been responsible for the observed congruency effect. Hence, this study demonstrates that sensorimotor experience with novel word referents is *sufficient* to establish such an effect, and to ground concepts in experience. However, it does not follow from this study that such direct experience is actually *necessary* in order to produce the effects, and therefore to achieve grounding of the novel words. To investigate exactly this issue is the aim of the present study.

It has sometimes been argued in the literature that *all* cognition is grounded in perceptual, motor, and emotional processes (Glenberg, 2015a), or that all mental
representations are experiential (Zwaan & Madden, 2005), and therefore that all mental symbols must be grounded in experience. This is illustrated by the claim that, “given that embodied constructs can explain cognition and action, there is no need to invoke abstract symbols” (Glenberg, 2015a, p. 169).

However, when accepting such a claim, one has to be very careful with what is actually meant by the proposition that a given symbol is grounded. For example, think of a person living in a landlocked place who has no direct experience with sharks actually swimming in his lower visual field. Let us further assume that he has made no perceptual experience with sharks at all - he somehow avoided watching TV shows and videos including sharks. All of the experience he has made with sharks is second hand information obtained through language, for example from books or stories. The information that sharks are associated with a downwards vertical position (they live in the water, under the surface of the ocean) therefore is part of his world knowledge, not his direct experience, and is obtained predominantly from linguistic sources of information. A very strong and simplistic version of the experiential trace model, which assumes that direct experience with a word’s referent is necessary in order to ground the respective concept, cannot account for congruency effects produced by such words. Hence, even if we only consider concrete nouns, direct experience alone cannot provide grounding for all possible words and the concepts denoted by them.

However, a closer look at the Chinese-Chinese dictionary argument offers one possible solution to resolve this issue. Imagine again the learner who struggles to learn Chinese with the monolingual Chinese-Chinese dictionary. Contrary to the former version of this argument, however, we now further assume that this dictionary has a small chapter at the beginning which translates some the most common symbols directly to the learner’s native language. With this version of the dictionary, she should now be able to figure out the meanings of some other symbols, and with those at hand the meaning of even more symbols unfold in a sort of chain reaction. Therefore, the Chinese-Chinese dictionary
argument does not imply that all symbols must be directly grounded; it also works if some (more specifically, a sufficiently large number) of symbols are - a grounded set of elementary symbols (Harnad, 1990). The meanings of other symbols can then be induced from these elements.

This solution to the argument was already suggested in the original article by Harnad (1990). Harnad proposed the mental symbol system to be a hybrid non-symbolic/symbolic system, with some elementary symbols being grounded in non-symbolic representations, to resolve the symbol grounding problem. Harnad suggested that, given the symbols “horse” and “stripes” are known and grounded, the meaning of “zebra” can easily be induced if we learn that a “zebra” is a “horse with stripes”. Hence, “ [...] because "horse" and "stripes" are grounded in their respective iconic and categorical representations, "zebra" inherits the grounding, through its grounded symbolic representation. In principle, someone who has never seen a zebra (but has seen and learned to identify horses and stripes) could identify a zebra on first acquaintance armed with this symbolic representation alone.” (Harnad, 1990, p. 343). Zwaan and Madden (2005) further specify this argument within the framework of the experiential trace model: According to their account, encountering the linguistic input “horse” and “stripes” activates the relevant experiential traces for their referents, which are then mentally combined. The new concept is then stored as a new, “constructed” experiential trace, which is in turn associated with the linguistic experiential trace “zebra”. Therefore, in the experiential trace model, co-occurrence of linguistic and perceptual input is not the only factor involved in establishing representations for concepts. The other two main factors are co-occurrences of perceptual input, and, importantly, co-occurrences of linguistic input (see the left panel of Figure 1).

Apart from solving the issue of how concepts for which we have no direct experience with the referents can be grounded, such an approach also provides a solution to another problem, namely scalability. It has been argued that school children learn 10-15 new words
and the corresponding concepts everyday (Landauer & Dumais, 1997), without necessarily being directly exposed to any of their referents, such as *Mesopotamia* or *uranium*. In fact, it has been estimated that average high school students know between 60,000 and 90,000 different words by the time they graduate, while they only encounter around 20,000 outside school and are taught around 200 per year (R. C. Anderson & Freebody, 1981; Nagy, Herman, & Anderson, 1985; Sternberg, 1987). The remaining word meanings have to be inferred from reading, that is, linguistic context (Landauer, Kireyev, & Panaccione, 2011). Therefore, conceptual knowledge can scale up rapidly, even if the direct experience with the world stays more or less the same from day to day. This scalability and increase in knowledge is thus most likely achieved by other mechanisms, such as extrapolating from linguistic input (Landauer & Dumais, 1997).

The aim of the present study is to demonstrate empirically that concepts can be grounded indirectly, through other symbols that are already grounded, and therefore can inherit their grounding through co-occurrence with other linguistic input. This assumption is sketched in the right panel of Figure 1. To be more precise, we hypothesize that upon encountering a novel word that was learned in a specific linguistic context, which in our study is constituted by words implying a vertical orientation, experiential traces associated to this context should become activated automatically, with the novel word serving as a cue to these experiential traces. This re-activation should then be observable in action congruency effects (Lachmair et al., 2011). This hypothesis compatible with the *language and situated simulation* (LASS) model (Barsalou, Santos, Simmons, & Wilson, 2008): According to the LASS model, perceiving a word immediately initiates linguistic processing of the stimulus, which includes activations of associated linguistic forms - in our case, words implying a vertical orientation. This activation is then followed by the activation of simulations (i.e., experiential traces) associated to the perceived word on the one hand, but also the activated linguistic forms on the other hand. Therefore, we argue that direct experience with a word’s referent is sufficient to establish grounding in sensorimotor
experience (Öttl et al., 2017), but it is not necessary for every word.

As mechanisms through with such an inheriting of grounding can be established, we propose that learning processes as discussed in the literature on abstract symbol systems (e.g. Collins & Quillian, 1969; Landauer & Dumais, 1997; Lund & Burgess, 1996; Smith & Medin, 1981) take place. In the experiments presented in this article, we will set up different learning scenarios in order to evoke learning mechanisms that are usually assumed to form relations between words in theories on abstract symbols: associative or distributional learning (Experiment 1 and 2), and symbolic learning (Experiment 3 and 4).

**Experiment 1**

A very simple mechanism through which such an indirect grounding could be established - and that is assumed to be a basic underlying process in associative networks and distributional semantic spaces - is associative or distributional learning. In distributional models of semantics (Landauer & Dumais, 1997; Lenci, 2008; Lund & Burgess, 1996), the meaning of a word is defined by the linguistic co-occurrences of words. The same assumption forms the basis of models of classical conditioning (Hebb, 1949; Rescorla & Wagner, 1972), where the association strength between two stimuli is a function of their co-occurrence pattern. Hence, a word such as “shark” can be understood by the previously mentioned landlocked person because it regularly occurs in the context of already known words, such as “ocean”, “fish”, and “swim”, and will therefore be related to the corresponding concepts.

In a series of experiments, Ouyang, Boroditsky, and Frank (2017) have demonstrated that humans engage in distributional learning, and that category memberships of novel words can be learned via this mechanism. To this end, they presented participants with novel words, organized in two distinct sets. One set of novel words was paired with different real words describing animals, and the other with different words describing vehicles. Each novel word was presented multiple times with each of its
associated real words during the learning phase. When participants then had to choose a referent for the novel word from a set of two pictures (one showing an animal, the other a vehicle, with both corresponding words not appearing in the learning phase), they had a higher probability of choosing a category member from the category the novel word was presented with.

In our first experiment, we employed a learning phase very similar to Ouyang et al. (2017), in which novel words were presented to participants, paired with real words from two different categories: Upwards-related and downwards-related words (Lachmair et al., 2011). Afterwards, a testing phase as used by Öttl et al. (2017) was employed, in which participants had to react towards these novel words with upwards or downwards hand movements. We assumed that participants learn to categorize the novel words as either upwards- or downwards-related words. This should be reflected in the participant’s answers when asked explicitly to assign these words to one of the two categories after the experiment: If a word was learned in the context of upwards-related words, there should be a higher probability that participants categorize it as upwards-related as opposed to downwards-related, and vice versa if learned in the context of downwards-related words.

We further assumed that the novel words will be indirectly grounded via the experiential traces of the respective real words they were presented with. These experiential traces should then be re-activated upon encountering the novel words. Therefore, our second hypothesis was that we observe action congruency effects (Lachmair et al., 2011; Öttl et al., 2017): Upwards-reactions should be faster for novel words that were presented with upwards-related words than for those paired with downwards-related words, and vice versa.

Method

Participants.
**Power analysis.** In order to identify an adequate sample size for our experiments, we performed a power analysis based on the data by Öttl et al. (2017), since we used an identical test phase, as well as a learning phase in which novel words were learned. Based on the parameters estimated from this data, we simulated $n$ new data sets (representing $n$ new participants) and performed the ANOVA analysis reported by Öttl et al. on this simulated data. This process was repeated 1,000 times for each value of $n$. Since the data was simulated from a model in which an interaction was present, the proportion of analyses detecting an interaction effect is an estimate of the test power.

For parameter values of $n \geq 42$, we obtained test power estimates of $power \geq .90$. Test power estimates were higher when employing the Linear Mixed Effect Model approach that will be used in the present study (see Baayen, Davidson, & Bates, 2008; Barr, Levy, Scheepers, & Tily, 2013). Analyzing Experiment 2 from Lachmair et al. (2011), which first introduced the Stroop-like task used by Öttl et al. (2017), gave very similar results, with test power estimates of $power \geq .90$ for sample sizes from $n \geq 38$. Following this power analysis, we set the planned sample size for our experiments to 45 participants.

**Participants.** All participants volunteered to participate in this experiment and gave informed consent. We tested 45 native German speakers (36 female, all right-handed, $M_{Age} = 24.62$ years, $SD_{Age} = 4.01$ years) whose data sets were passed to the data analysis. Data sets from 7 additional participants had to be excluded due to technical problems in the data collection (2 participants) or due to error rates higher than 10% in at least one experimental condition (exclusion criteria taken from Lachmair et al., 2011), which affected 5 participants. Participants received either money or course credit for their participation.

**Material.**

**Novel words.** As in the study by Öttl et al. (2017), we used 8 different novel words. To select these novel words, we handed a questionnaire with 30 different novel words, which had a German phonemic and subsyllabic structure and were constructed using Wuggy (Keuleers & Brysbaert, 2010), to 60 native German-speaking participants (39
female, $M_{\text{age}} = 23.8$ years, $SD_{\text{age}} = 4.9$ years). Participants had to choose for each novel word whether they associated it with an upwards- or a downwards-location, and to guess in cases where they felt they could not give a sensible answer. Of the 30 novel words, we selected the eight word for which the proportion of up-ratings (and therefore also the proportion of down-ratings) was closest to .5. The selected novel words with their respective rating proportions can be seen in Table 1. It was counter-checked that these novel words were not accidentally rare real words.

**Real Words.** We further selected eight real words (four up-words, four down-words) the non-words would be paired with, to obtain a design for our learning phases similar to Ouyang et al. (2017), with the difference that we used four instead of three words for each category. These real words were obtained from the item material used in the study by Lachmair et al. (2011). Out of the 39 up-words and 39 down-words employed in this study, we explicitly selected those words that showed the strongest and most consistent action-congruency effects in Lachmair et al.’s experiments (see Table 1).

To ensure that the real words we selected for this study indeed evoke these congruency effects on their own, without the context of other up- or down-related words, we conducted a pre-test study with 12 native German speaking participants (10 female, 12 right-handed, $M_{\text{age}} = 25.3$ years, $SD_{\text{age}} = 4.1$ years) who did not participate in any other experiment reported in this article. This experiment was identical to the test phase in Öttl et al. (2017) and the test phase employed in the present experiment, and will be described in full detail in the following sections. The only difference between the pre-test and the actual experiments was that the eight real words in Table 1 were used instead of novel words. We observed a significant location $\times$ response direction congruency effect on reaction times in the pre-test study: A Linear Mixed Effects Model (LMM) including this interaction effect explained the data significantly better as compared to a model including only two main effects ($\chi^2(1) = 4.08, p = .043$). The mean reaction times per condition can

---

2 One novel word, “Mende”, is a very rare German surname.

3 Both models also included random intercepts for subjects and items, as well as by-subject random slopes.
be seen in Figure 2.

**Stimuli and Apparatus.** The actual stimuli presented to a participant in the learning phase were short phrases, always consisting of a novel word and a real word which were connected by the word “und” (“and”), as in the study by Ouyang et al. (2017). Each novel word was paired either with all four up-words, or with all four down-words. Four of the novel words were paired with up-words, and the remaining four were paired with down-words. These pairings were fixed within each participant, but randomized over participants. Two important facts follow from this assignment strategy: First, since each participant only encounters each of the novel words with either the four up-words or the four down-words, models of associative or distributional learning would predict that the associative strength between a novel word and the words it is not presented with is very low (or zero), but higher for the words it is presented with. Secondly, since the assignment of each novel word to either the up-words or the down-words is randomized between participants, potential idiosyncratic effects for the novel words (which also are controlled for in our pre-test rating study) should not systematically affect our results.

In half of the learning phase trials, the real word was the first word in a pair, while the novel word was the first word in the other half. Thus, 64 different stimuli were presented to each participant in the learning phase, 4 novel words × 4 real words × 2 categories (up vs. down) × 2 orders (real word first vs. novel word first). In the test phase, only the novel words were presented to the participants.

In addition to a visual presentation, all phrases were also played as auditive stimuli in the learning phase. This was done to ensure that participants indeed paid attention to the stimuli - since they had to perform no specific task in this phase - and to make the learning phase slightly more interesting. The auditive stimuli had a duration of about 2 seconds each, with short intervals after each word in the phrase, and were recorded by a
male native German speaker.

Responses in the test phase were recorded using a vertically mounted standard computer keyboard. This keyboard had a specifically constructed overlay with only four, relatively large buttons: A lower button, two middle buttons (a lower and an upper middle button), and an upper button (see Figure 3). The experiment was conducted on a standard computer; the learning phase was programmed using PsychoPy (?), the test phase using Psychtoolbox for Matlab (Brainard, 1997).

Additionally, participants had to perform an explicit judgement task, where they were asked to indicate for each of the eight novel words if they associated the respective word with an upwards- or a downwards-location (similar to the questionnaires used in the novel word pre-test ratings). They were asked to guess in cases where they felt they could not make a clear decision.

**Procedure.** The experimental procedure consisted of three parts: A learning phase, a test phase, and an explicit judgement task. After each of these parts, the experimenter started the next part (or ended the experiment, after the explicit judgement task). The whole experiment took about an hour to finish.

**Learning phase.** The learning phase was similar to the learning phase introduced in Ouyang et al. (2017). Participants were instructed that they would be presented with pairs of words, and were asked to pay attention to these pairs. We did not ask them to explicitly learn the word combinations they would be presented with, and did not to include any task that would force participants to pay attention to the stimuli presented to them.

In each learning trial, a phrase was presented in white letters on a black screen for 2500 ms. While the phrase was on-screen, the same phrase was also played to the participants as an auditive stimulus from a single loudspeaker located in front of the participants. The stimulus presentation therefore differed from Ouyang et al. (2017), in that we simultaneously presented visual and auditive stimuli instead of auditive stimuli
only. After the visual stimulus disappeared, participants could start a new trial by pressing the space bar. A new trial then started after a 500-ms delay.

Each learning phase block consisted of all 64 different stimulus pairs selected for the participant, presented in random order. The whole learning phase consisted of 7 experimental blocks. Thus, each word pair was presented 14 times to the participant during the course of the learning phase, 7 times with the novel word being the first word, 7 times with the real word being the first word. Participants were not informed when a new block started, but were free to pause whenever they wanted between trials. The learning phase took about 30 minutes to finish.

**Test phase.** The test phase was identical to the one used by Öttl et al. (2017). At the beginning of each trial, participants had to press the two middle buttons of the keyboard with one hand each. Whether the dominant hand had to press the upper or the lower middle button was counterbalanced between participants. Each trial started with a fixation cross which was presented in the centre of the screen for 750 ms. Then, one of the novel words was presented in the centre of a white screen in one of four colours (blue, green, red, or orange). Participants were instructed according to the *colours* of the words: For two of the colours, they had to react with an upwards movement (release the upper middle button and press the upper button with the same hand); for the remaining two colours, they had to react with a downwards movement (release the lower middle button and press the lower button with the same hand). Two colours were mapped to each response direction, as in the original experiments (Lachmair et al., 2011; Öttl et al., 2017), to increase the difficulty of the task and therefore to engage participants in slightly longer processing of the stimuli before making their response. The assignment of response directions to colours was counterbalanced across participants. The time participants took to release the respective middle button was recorded as the reaction time (Lachmair et al., 2011; Öttl et al., 2017). The time interval between releasing the middle button and pressing the outer button will be reported as movement time. The novel word remained
on-screen until one of the middle buttons was released, or disappeared after a fixed duration of 1500 ms. In the latter case, the feedback “too slow” was presented to the participants. If the wrong key was released or pressed, the feedback “Error” was presented when participants pressed the outer button. After each trial, participants were instructed to press the two middle buttons again, which started the next trial after a 1000-ms delay.

In each experimental block of the test phase, each novel word was presented once in each of the colours; a block therefore consisted of 32 trials (8 novel words $\times$ 4 colours). Participants had the opportunity to get used to the response apparatus and the task in a practise block consisting of 16 trials. In this practise block, two different letter strings (XXXX and YYYY) were presented to the participants two times in each of the in the four colours. The whole test phase consisted of 8 experimental blocks, and took about 20 minutes to finish.

Explicit Judgement Task. After the test phase, participants performed the explicit judgement task, which took about one or two minutes to finish.

Design. The present experiment deployed a $2 \times 2$ design. These factors were learning context for the novel words (up-words vs. down-words) and response direction (upwards vs. downwards). Both factors are manipulated within subjects and within items ($items$ being the eight novel words).

Results

Since it was randomly determined for each participants which novel words would be presented in which learning context, the distribution of novel words over these contexts is not completely even. The distribution can be seen in Table 2.

Test Phase Results. Trials in which an error was made (2.2 %), as well as trials with reaction times below 100 ms (0 %) were excluded from the data analysis (as in Lachmair et al., 2011). Mean reaction times by learning context and response direction can be seen in Figure 4.
In all the experiments reported here, we analysed the logarithmic test phase reaction times (Baayen & Milin, 2010) by employing LMEMs, using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) for R (R Core Team, 2017). Since our hypothesis predicts an interaction between learning context and response direction, we compared two models by performing a likelihood-ratio test: The baseline model contained fixed effects for learning context and response direction, random intercepts for participants and items, by-participant random slopes for both fixed effects and their interaction, as well as by-item random slopes for both fixed effects (Barr et al., 2013). We then tested whether a model that additionally contained a fixed interaction effect between learning context and response direction significantly improved the model. This was not the case; the interaction effect did not improve the model ($\chi^2(1) = 0.001, p = .973, \beta = -0.00, t = -0.03$). In an analysis of logarithmic movement times (defined as the time between releasing a middle button and pressing the upper or lower button), including an interaction effect did also not improve the baseline model ($\chi^2(1) = 0.770, p = .380, \beta = -0.011, t = -0.88$).

In this and all of the following experiments, we obtained very similar results when we restricted the analysis only to items for which the participants’ answer in the explicit judgement task and the actual learning context coincided. Furthermore, in this and all of the following experiments, we did not find an interaction effect for the accuracy data in an GLMEM analysis.

**Explicit Judgement Task Results.** The proportions of participants’ responses by learning context can be seen in Figure 4. We set up a GLMEM using lme4 (Bates et al., 2015) to perform a logistic regression analysis. This model predicted whether or not a novel word was being rated as upwards-located, and contained random intercepts as well as random slopes for the learning context for both participants and items. We performed a likelihood-ratio test to compare two models, one that additionally contained a fixed effect

---

4A by-item random slope for the interactions could not be included since the models did not converge in this case

5Model parameters for reaction times analyses for all experiments can be found in the Appendix
for learning context and one that did not. The model which contained a learning context
fixed effect predicted the participants’ answers significantly better ($\chi^2(1) = 12.01, p < .001,$
$\beta = -1.63, z = -4.32,$ ). In Figure 4 and the following figures showing explicit judgement
task results, the .95 confidence intervals are obtained from the model containing a learning
context fixed effect, and are computed using the effects package for R (Fox, 2003).

To test whether ratings in both learning contexts differed from chance level (a
proportion of up-ratings of .5), we set up two GLMEMs, predicting the ratings with an
intercept as well as random intercepts for both participants and items within each learning
context, respectively. If the ratings were at chance level, one would obtain an intercept of
zero in these models; intercepts which differ significantly from zero indicate a proportion of
up-ratings above (positive intercepts) or below (negative intercepts) chance level. For the
up-words learning context, we obtained a significant positive intercept ($\beta = 0.70, z = 3.37,$
$p = .001$). For the down-words learning context, we obtained a significant negative
intercept ($\beta = -0.83, z = -4.02, p < .001$).

Discussion

In this experiment, we presented novel words paired with either upwards- or
downwards-related words in order to evoke associative learning. We then tested whether
these novel words then elicited an action-congruency effect in a behavioural task as used by
Lachmair et al. (2011), and whether participants categorize the novel words according to
their respective learning contexts (Ouyang et al., 2017).

The results of this experiment were mixed: In the explicit categorization task,
participants showed a clear tendency towards assigning a novel word to the category it was
presented with. This extends the findings of the original study by Ouyang et al. (2017),
where very concrete basic-level categories were used, whose category members typically
share many features. The category members used as stimulus material in the present
experiment arguably share far less features, apart from being upwards- or
downwards-related. Still, participants assigned the novel words more frequently to the
category they were presented with. However, we did not find any effect of the learning
context on reaction times in the test phase, that is in the behavioural task we employed.
Thus, contrary to our hypothesis, we did not observe any signs for the re-activation of
experiential traces in this first experiment.

A possible explanation for this pattern relies on the fact that participants are forced
to make an explicit judgement in the categorization task, while the behavioural test phase
is a far more implicit task, comparable for example to implicit association tasks. Therefore,
it is quite possible that vertical locations plays no role in the participants’ representation of
the novel words, unless the task explicitly requires a judgement about this feature. This
would render the results in the categorization task as a purely retrieval-based phenomenon,
where the learned vertical location is not automatically activated for the novel words, but
only if such an activation is necessary. To put it in other words, participants might only
activate the vertical location associated with the novel words when they are asked to do so.
This explanation runs contrary to our hypothesis that an automatic activation of
experiential traces can be inherited through co-occurrence of linguistic input alone.

Another problem with the learning phase employed in this study is that potential
referents and representations for the novel words are highly underdetermined, basically on
the level of “something more or less located in upwards positions”. The representation of
such a concept most likely is very unclear. This is a stark contrast to the experiment by
Öttl et al. (2017), where an action-congruency effect was observed: The novel words in this
experiment had very clear referents (concrete object that were shown to the participants),
and therefore most likely a very clear representation. In our second experiment, we tried to
address this issue by setting up a different learning phase which should result in a much
clearer representation of the concepts denoted by the novel words.
Experiment 2

In Experiment 2, we set up a different learning phase than in Experiment 1, while using the same test phase and the same explicit judgement task. The learning phase now consisted of real natural sentences collected from text corpora, in which either upwards- or downwards related words were replaced with novel words (Lazaridou, Marelli, & Baroni, 2017).

The idea that a word’s meaning is determined by the natural linguistic context it occurs in is very prominently expressed in the distributional hypothesis (Harris, 1954; Sahlgren, 2008), which is the theoretical foundation for distributional models of semantics (Landauer & Dumais, 1997; Lund & Burgess, 1996). The learning phase we employed in Experiment 2 directly implements - on a smaller scale - the distributional learning mechanism these models rely on.

As already mentioned, we replaced real words in natural text with novel words in order to create the material for this experiment. More specifically, we made use of textual chimeras (Lazaridou et al., 2017), where two different, but related, real words were replaced with a single novel word. For example, if Seige replaced both Mond (moon) and Wolke (cloud), it will most likely describe a large object far up in the sky. For the purpose of our study, this technique has the advantage that we still can exclude that a participant has made direct experience with the novel words’ referents, since these referents are just mashed-up chimera concepts.

Our hypotheses for Experiment 2 were the same as for Experiment 1 - we expected to observe action congruency effects for the novel words, for the same reasons as discussed there, and we expected to replicate the findings of the categorization task in Experiment 1.

Method

Participants. All participants volunteered to participate in this experiment and gave informed consent. We tested 45 native German speakers (33 female, all right-handed, $M_{Age} = 24.82$ years, $SD_{Age} = 4.48$ years) whose data sets were passed to the data analysis.
Data sets from 4 additional participants had to be excluded due to high error rates. Participants received either money or course credit for their participation. None of the participants of Experiment 2 had already participated in Experiment 1.

Material. As novel words, we used the same novel words as we did in Experiment 1. As real words to be replaced in the sentences for this material, we selected 16 words: The eight real words already used in Experiment 1, four additional up-words, and four additional down-words taken from Lachmair et al. (2011). Each real word from Experiment 1 was paired with one of the additional words with the same vertical position in order to build up a chimera, as can be seen in Table 3. It was taken care that both words of a chimera pair denoted related concepts, in order to create coherent chimeras.

For each real word, we selected ten different (German) sentences from the Wortschatz Leipzig corpus (http://wortschatz.uni-leipzig.de/) and the DWDS Kernkorpus 20 (http://www.dwds.de/ressourcen/kernkorpus/), which contained the word in question exactly once in its singular form, with no restrictions on the word’s position within the sentence. The sentences had a length between 7 and 29 words. In these sentences, we then replaced the two words for each chimera pair with one of the novel words (for example, both Wolke and Mond could be replaced with Seige for some participants), resulting in 20 sentences for each novel word. It was taken care that these 20 sentences were compatible with one another and did not contain contradictory information. The assignment of novel words to the chimeras was randomized across participants (for example, for some other participants Keller and Graben could be replaced with Seige). All novel words were treated as feminine nouns and all sentences were adjusted accordingly, since this sounded most natural for the novel words we used.

Unlike in Experiment 1, we did not use auditive stimuli in Experiment 2, since we considered the task of reading the sentences as being interesting enough for the participants. The test phase and explicit judgement task in Experiment 2 were identical to Experiment 1, and therefore we used the same material for these phases.
**Procedure.** Identical to Experiment 1, Experiment 2 consisted of three phases: A learning phase, a test phase, and an explicit judgement task. The test phase and the explicit judgement task were identical to Experiment 1.

For the learning phase, participants received no further instruction but to carefully read the sentences. The learning phase consisted of two sub-phases: A blocked presentation of the sentences, followed by a completely randomized presentation of the same sentences. In the blocked presentation, all 20 sentences containing the same novel word were presented in blocks. The order of the sentences within a block randomized, as was the order of blocks. The purpose of this blocked phase was to provide the participants with enough continuous experience with every novel word to establish a fairly clear representation for it. In the subsequent completely randomized presentation, all 160 sentences were presented in a completely randomized order, without any blocks, to ensure that participants had recent experience with each word at the end of the learning phase.

Each sentence was presented to the participants in white letters on a black screen, and most sentences were presented across several lines on the screen. Each sentence remained on-screen for a minimum of 4000 ms, without a maximum presentation time. In order to initiate the presentation of the next sentence, participants had to press the space bar on their keyboard, upon which the sentence disappeared from the screen and a new sentence was presented after a 500-ms-delay.

The learning phase therefore had a minimum length of 320 sentences * 4500 ms = 24 minutes. However, no participant took less than 29 minutes to complete the learning phase ($M = 37$ minutes), which indicates that the participants actually performed the task of reading the sentences.

**Results**

The distribution of novel words over the learning contexts (Up-context vs. Down-context) can be seen in Table 2
**Test Phase Results.** Erroneous trials (2.4 %), as well as trials with reaction times below 100 ms (< 0.01 %) were excluded from the data analysis. Mean reaction times by learning context and response direction can be seen in Figure 5. The LMEM analysis we conducted was identical to Experiment 1. Including a fixed interaction effect did not improve the mixed-effect model for logarithmic reaction times ($\chi^2(1) = 0.137, p = .711, \beta = -0.043, t = -0.37$). The same pattern was found in an analysis of logarithmic movement times ($\chi^2(1) = 1.594, p = .207, \beta = 0.014, t = 1.26$).

**Explicit Judgement Task Results.** The proportions of participants’ responses by learning context can be seen in Figure 5. A GLMEM containing a learning context fixed effect predicted the participants’ answers significantly better ($\chi^2(1) = 24.81, p < .001, \beta = -3.73, z = -5.75$).

In the up-words learning context, we obtained a significant positive intercept ($\beta = 1.44, z = 5.36, p < .001$), indicating that the number of Up-Ratings was above chance level. For the down-words learning context, we obtained a significant negative intercept ($\beta = -1.85, z = -4.63, p < .001$), indicating that the number of Up-Ratings was below chance level.

**Discussion**

In Experiment 2, we employed a similar paradigm as in Experiment 1, with a different learning phase. Participants read sentences containing novel words, which replaced two real words with the same vertical orientation. The learning phase of Experiment 2 is therefore a rather naturalistic setting for the acquisition of new words: We encounter them in natural text, and after doing so several times, we have a rough idea of what their meaning is.

However, the pattern of results we obtained in Experiment 2 was identical to Experiment 1. We again found a clear dissociation between behavioural results in the test phase on the one hand, and explicit judgement task results on the other hand. When asked to judge the vertical orientation of the novel words, participants were very much able to do...
this. This is an interesting result by itself, as it demonstrates that learning about features of a novel concept can be achieved by presenting the respective novel words in the context of natural sentences, thereby extending the findings by (Ouyang et al., 2017) to another, more naturalistic learning scenario. Remarkably, this learning occurs even if the novel words do not directly correspond to an actual concept with which the learner already has experience, but are instead textual chimeras created from two category members (Lazaridou et al., 2017). This result supports the distributional hypothesis that word meanings are at least partially determined by the linguistic context the words occur in (Harris, 1954; Landauer & Dumais, 1997; Lund & Burgess, 1996; Sahlgren, 2008). At least, semantic features of novel words can be induced from such contexts, even with rather small amounts of learning experience for each novel word (Lazaridou et al., 2017).

However, the results of the test phase do not indicate that there is any automatic activation of experiential traces that influenced the upwards- and downwards-movements with which participants reacted according to the words, contrary to our initial hypothesis. However, the issue of potentially unclear representations as discussed for Experiment 1 still remains, albeit to a far lesser degree. We can reasonably assume that each of the 20 sentences containing a novel words restricts the set of possible referents to the word in question; however, this does not imply any definite information about the size of the final set of possible referents. In order to address this issue more thoroughly, we implemented different learning phases in Experiment 3 and 4, where we explicitly restrict the referent set for the novel words to one specific concept. By doing this, we tried to make our learning phases more comparable to the study by Öttl et al. (2017), where the novel words also referred to one specific concept.

**Experiment 3**

In Experiment 3, we set up a new learning phase where participants explicitly learned that each novel word refers to exactly one real word, and therefore to a definite single concept.
To this end, novel words were paired with exactly one real word each. Participants were instructed to learn the pairings, and also had to recall all pairings correctly in order to complete the learning phase. It was thereby made obvious that the novel words were to be seen as new labels for the concepts denoted by the real words.

In this, the learning phase of Experiment 3 is highly similar to explicitly learning the vocabulary of a second language (L2) from a dictionary, where learners have a previously unknown word (the novel word), and have to learn which word in their native language this word corresponds to. Interestingly, it has already been shown that the results by Lachmair et al. (2011) can be replicated with L2 speakers responding to L2 language material (Dudschig, de la Vega, & Kaup, 2014). However, in such a study, one cannot control which kind of experience the participants have with their respective L2; after all, if they sometimes use their L2 in real-life situations, they do encode sensorimotor experience along with L2 language experience. Therefore, in order to ensure that participants only have specific linguistic experience with the novel words they are tested on, we have to ensure that they encounter them for the first time in a controlled experimental setting.

Method

Participants. All participants volunteered to participate in this experiment and gave informed consent. We tested 45 native German speakers (32 female, all right-handed, $M_{\text{Age}} = 23.18$ years, $SD_{\text{Age}} = 3.98$ years) whose data sets were passed to the data analysis. Data sets from 4 additional participants had to be excluded due to high error rates, as described earlier, and data sets from 12 additional participants had to be excluded due to technical errors and incomplete data. Participants received either money or course credit for their participation. None of the participants of Experiment 3 had already participated in Experiment 1 or 2.

Material. As novel and real words, we used the same words as described in Experiment 1. For the learning phase, each of the novel words was randomly paired with
exactly one of the real words, resulting in eight word pairs. Word pairs were created randomly for each individual participant. The words in a pair were always separated by an equals sign, with the novel word always being on the left side and the real word always being on the right side of the equals sign. Hence, participants were presented with stimuli such as \textit{Essede} = \textit{Sumpf} (\textit{Essede} = \textit{swamp}). The test phase and the explicit judgement task in Experiment 3 were identical to Experiment 1 and 2.

\textbf{Procedure.} Experiment 3 again consisted of three different phases: A learning phase, as well a test phase and an explicit judgement task for which the procedure was exactly as described for the previous experiments.

At the beginning of the learning phase, participants were instructed that they were to be presented with word pairs consisting of a real and a novel word. They were further instructed to learn these pairings. After participants read the instruction, a presentation phase started where the eight word pairs were presented to the participants one after another. Each word pair remained on the screen until participants decided to see the next word pair by pressing the space bar. The order in which the word pairs appeared was randomized for each learning phase.

After presentation of the eight word pairs, a recall phase started, in which participants were presented with the novel words and instructed to recall the respective real words they were paired with in the presentation phase. Novel words were presented one at a time, in the upper half of the screen. Participants were instructed to type in the respective real word using the keyboard. Changes to the input could be made by using the backspace key. After participants confirmed their input by pressing the enter key, they were provided with feedback - either \textit{Korrekt!} (\textit{Correct!}), or \textit{Fehler!} (\textit{Error!}) - which remained on-screen for 1,000 ms.

After providing recall answers for all eight novel words this way, participants received feedback on how many of these answers were incorrect. If this number was zero, the learning phase was completed. If this number was greater than zero, the learning phase
(both presentation of word pairs and recall phase) was repeated, with the same word
pairings, but with a new randomized item order (for both presentation and recall phase).
Each participants had to repeat learning phases until she/he made exactly zero recall
errors in one learning phase. With this procedure, we wanted to ensure that participants
indeed correctly memorized the pairings for all eight novel words.

Results

The distribution of novel words over the learning contexts (Up-context vs. Down-context)
can be seen in Table 2

Test Phase Results. As for the previous experiments, trials in which an error
was made (2.6 %), as well as trials with reaction times below 100 ms (< 0.01 %) were
excluded from the analysis. Mean reaction times by learning context and response
direction are displayed in Figure 6.

In the LMEM analysis of logarithmic reaction times as described in the previous
experiments, the model including an interaction effect between learning context and
response directions did not perform significantly better than a model without this effect
($\chi^2(1) = 2.712, p = .100, \beta = 0.017, t = 1.68$). This was also the case in an analysis of
logarithmic movement times ($\chi^2(1) = 5.321, p = .053, \beta = 0.037, t = 1.81$).

Explicit Judgement Task Results. The proportions of participants’ responses
by learning context can be seen in Figure 6. The explicit judgement task data was
analysed using the same GLMEM approach as described for the previous experiments.
Including a fixed effect for learning context significantly improved the baseline model
($\chi^2(1) = 26.76, p < .001, \beta = -4.60, z = -5.81$). Furthermore, when setting up models for
both learning contexts separately, we found a significant positive intercept for words
learned in the up-context ($\beta = 2.02, z = 4.94, p < .001$), as well as a significant negative

\footnote{Although this effect in the logarithmic movement times is near significance level, it runs in the opposite
direction as expected in our hypothesis: Upwards movements are numerically faster for words learned in a
Down-Context, and vice versa.}
intercept for words learned in the down-context ($\beta = -2.08$, $z = -5.27$, $p < .001$), indicating that both groups differed from chance level in the expected direction.

Discussion

We observed the same pattern of results as for the previous experiments, namely a dissociation between the behavioural data, where we did not find a congruency effect between learning context and response direction, and the explicit judgement task data, where we found that participants were very capable of making correct explicit judgements about the novel words’ vertical locations.

Initially, these results are not in line with the findings by Dudschig et al. (2014), that regular L2 words can elicit action congruency effects. However, there are important differences between the novel words in the current experiment and regular L2 words. Typical L2 language users arguably have far more experience with L2 words than our participants had with the novel words, in at least three ways: They have been more heavily exposed to the L2 words, both in terms of frequency and over a longer period; they had the opportunity to actually ground the L2 words in sensorimotor experience; and they have actually used and uttered the L2 words in communicative settings.

Another issue is that the novel words in Experiment 3 do, strictly speaking, not serve any communicative purpose, since are only new labels for concepts that already have a label. Therefore, using or applying these new labels would only increase the communicative effort without bringing extra benefit - unlike regular L2 words, that enable a speaker to address a new language community. In order to address this issue, we set up a fourth experiment in which novel words were learned as labels for novel concepts, which renders them informative and potentially useful for communicative purposes.

Experiment 4

In Experiment 4, participants learned the novel words as labels for novel concepts. In fact, learning new words makes most sense if they are labels for concepts or ideas that do not
yet have a label, so these can more easily be referred to, thereby lowering the communicative effort.

Once we have learned a sufficient amount of basic concepts, most new concepts we acquire are the result of conceptual combination (Murphy, 2002b; Ran & Duimering, 2009), that is, a product of bringing together and establishing a relation between two familiar concepts to form a new concept. This is prominent in the observation that many new inventions and discoveries are labelled using compounds, such as spaceship, computer virus, or information warfare (Thagard, 1984).

The reasoning that conceptual combination is an important way of acquiring new concepts is in fact a central component of the proposals by Harnad (1990) concerning the symbol grounding problem. As discussed in the Introduction to this article, Harnad argues that a novel concept such as zebra can be grounded by learning that zebra = horse + stripes, if the basic concepts horse and stripes are grounded. Zebra therefore poses a perfect example of a new concept that is formed by combining two familiar concepts, and that is labelled with a novel word. If we follow this argumentation, we can hypothesize by keeping our initial assumptions that the concept zebra should then be indirectly grounded via horse and stripes. If then one has a combined novel concept for which the basic concepts elicit action congruency effects, we can hypothesize that the novel concept inherits these congruency effects.

Method

Participants. All participants volunteered to participate in this experiment and gave informed consent. We tested 45 native German speakers (36 female, all right-handed, $M_{Age} = 22.44$ years, $SD_{Age} = 3.53$ years) whose data sets were passed to the data analysis. Data sets from three additional participants had to be excluded due to high error rates, as described earlier. Participants received either money or course credit for their participation. None of the participants of Experiment 4 had already participated in any of
the previous experiments reported here.

**Material and Procedure.** In Experiment 4, we employed a learning phase that was very similar to the one in Experiment 3, with the difference that each novel word was paired with a phrase describing a novel concept instead of a single noun describing an existing concept. These phrases were obtained as follows: First, we constructed four different phrases for each of the four up-words and four down-words used in the previous experiments, resulting in a total of 32 phrases. Each of these phrases consisted of the respective up- or down-word, followed by a preposition and another noun. The phrases used in this Experiment are shown in Table 4. The up- or down-words were always the head of the phrase, and determined the semantic category of the concepts described in the phrases (for example, a *roof made of silk* is an instance of a *roof*). This set of 32 phrases was then handed out to 30 native German speaking participants (25 female, $M_{Age} = 23.33$ years, $SD_{Age} = 4.23$ years), who did not participate in the actual experiment, in a rating questionnaire. Participants were then asked to indicate for each phrase on a 7-point rating scale (a) how familiar they were with the described concept (1 - unfamiliar, 7 - familiar), (b) how plausible they perceived the described concept to be, that is, how well they could imagine the concept to actually exist (1 - implausible, 7 - plausible), and (c) if the described concept would be located in a lower or an upper position in the real world (1 - lower, 7 - upper). We then selected exactly one phrase from each of the eight sets of four phrases including a specific up-word or down-word. The selection criteria were that the concepts described by the phrases were as unfamiliar as possible on the one hand, while on the other hand they were also clearly rated to be located in lower or upper positions\(^7\). Furthermore, given these criteria were sufficiently fulfilled, the concept should also be judged to have at least medium plausibility, with a preference for more plausible concepts. The final selection of phrases that was used in Experiment 4 can be seen in Table 4.

\(^7\)The mean location ratings for the selected novel concepts are comparable to the ones obtained for the original stimuli by Lachmair et al. (2011)
Apart from the fact that these phrases were paired with the novel words in the learning phase, instead of just the four up-words and the four down-words themselves, Experiment 4 was identical to Experiment 3 concerning both Material and Procedure.

Results

The distribution of novel words over the learning contexts (Up-context vs. Down-context) in Experiment 4 is shown in Table 2.

Test Phase Results. Again, erroneous trials (3.1 %), as well as trials for which reaction times were under 100 ms (0.0 %) were excluded from the analysis. Mean reaction times by learning context and response direction are displayed in Figure 7.

A Linear Mixed Effects Model including an interaction effect between learning context and response directions did not perform significantly better than a model without this effect ($\chi^2(1) = 1.094, p = .295, \beta = -0.010, t = -1.05$). Including an interaction effect did also not improve the baseline model in an analysis of movement times ($\chi^2(1) = 0.129, p = .719, \beta = -0.007, t = -0.36$).

Explicit Judgement Task Results. The proportions of participants’ responses by learning context can be seen in Figure 7. Including a fixed effect for learning context significantly improved the baseline GLMEM, as described in the previous experiments ($\chi^2(1) = 19.45, p < .001, \beta = -3.456, z = -5.465$). When setting up two separate models for both learning contexts, we found a significant positive intercept for words learned in the up-context ($\beta = 1.414, z = 5.15, p < .001$), as well as a significant negative intercept for words learned in the down-context ($\beta = -1.529, z = -5.47, p < .001$). Therefore, both groups differed from chance level in the expected direction.

Discussion

In Experiment 4, participants learned novel words as labels for concepts. Unlike in Experiment 3, where these concepts were existing and already labelled concepts such as *swamp*, participants in Experiment 4 learned the novel words as labels for novel (or...
unfamiliar) yet plausible concepts such as *swamp in a mine*. Therefore, learning labels for these concepts made sense from a communicative viewpoint, as they serve a communicative purpose - with the label at hand, one is able to talk efficiently about concepts that had to be described using a complex phrase before.

However, the pattern of results in Experiment 4 is still identical to the previous experiments: We did not find any congruency effect in the test phase reaction time data; on the other hand, results from the explicit judgement task data shows that participants were clearly able to infer the novel words' vertical position correctly from the learning phase. This especially shows that it apparently plays no role for our experimental paradigm whether novel words were learned as labels for existing, labelled concepts (Experiment 3) or novel, combined concepts that did not yet have a label (Experiment 4).

**Experiment 5**

In Experiments 1 to 4, we have found no evidence for a re-activation of sensorimotor experiential traces when participants were presented with novel words learned from purely linguistic input. However, re-considering our hypothesis of indirect grounding (see the Introduction and Figure 1) offers a possible explanation for this pattern of results: Novel words learned in a specific linguistic context – in our case, existing words – could serve as a cue to re-activate these existing words, which then in turn re-activate sensorimotor experiential traces. Assuming that no direct association is formed between the novel words and the sensorimotor traces, or that our learning phases were not sufficient to establish such an association, we could have a case of mediated activation at hand. In this case, re-activation of sensorimotor traces for novel words learned from purely linguistic input can be expected to take longer than for words directly associated with sensorimotor experience (as in Öttl et al., 2017), due to the necessity of an intermediate step. This would prevent action-congruency effects if a response is already being made before the relevant sensorimotor traces are activated. In Experiment 5, we examined this possibility by
introducing a delay between the presentation of the target word and the colour specifying the response direction, so that participants could process the word for a short amount of time before being making a response.

Participants

All participants volunteered to participate in this experiment and gave informed consent. We tested 45 native German speakers (37 female, 42 right-handed, $M_{Age} = 24.02$ years, $SD_{Age} = 5.59$ years) whose data sets were passed to the data analysis. Data sets from two additional participants had to be excluded due to high error rates, as described earlier. Participants received either money or course credit for their participation. None of the participants of Experiment 5 had already participated in any of the previous experiments reported here.

Material and Procedure

In most respects, Experiment 5 was a replication of Experiment 4: The learning phase and the explicit judgement phase of Experiment 5 were identical to those of Experiment 4. The test phase was almost completely identical to all the previous test phases. However, there was a difference in the trial procedure: When a novel word was presented on the screen, it first appeared in a grey colour. Only after a delay of 500 ms did the colour of the word change into one of the four colours specifying the correct response direction (blue, green, red, or orange).

Results

The distribution of novel words over the learning contexts (Up-context vs. Down-context) in Experiment 4 is shown in Table 2

Test Phase Results. Again, erroneous trials (2.3 %), as well as trials for which reaction times were under 100 ms (< 0.01 %) were excluded from the analysis. Mean reaction times by learning context and response direction are displayed in Figure 8.
With respect to logarithmic reaction times, a Linear Mixed Effects Model including an interaction effect between learning context and response directions did not perform significantly better than a model without this effect ($\chi^2(1) = 0.160, p = .690, \beta = 0.006, t = 0.83$). Including an interaction effect did also not improve the baseline model in an analysis of logarithmic movement times ($\chi^2(1) = 0.646, p = .422, \beta = -0.011, t = -0.81$).

**Explicit Judgement Task Results.** The proportions of participants’ responses by learning context can be seen in Figure 8. Including a fixed effect for learning context significantly improved the baseline GLMEM, as described in the previous experiments ($\chi^2(1) = 24.19, p < .001, \beta = -3.456, z = -7.19$). When setting up two separate models for both learning contexts, we found a significant positive intercept for words learned in the up-context ($\beta = 1.618, z = 6.14, p < .001$), as well as a significant negative intercept for words learned in the down-context ($\beta = -1.442, z = -6.68, p < .001$). Therefore, both groups differed from chance level in the expected direction.

**Discussion**

In Experiment 5, we introduced a brief delay between target word presentation and response, under the assumption that a mediated activation of experiential traces could take place that was not detected by the previous experiments. However, as in the previous experiments, we did not observe an action-congruency effect in Experiment 5. At the same time, participants were again very well capable of making explicit judgements about the vertical location associated to the novel words.

Hence, Experiment 5 further supports the interpretation of Experiments 1-4, that experiential traces are not activated during processing of the novel words. More specifically, we do not find evidence that experiential traces are activated in a mediated fashion, at least not with a 500 ms delay between presentation of the target and response.

---

8All random slopes for interactions had to be excluded, since models including these effects did not converge.
General Discussion

In the present study, we investigated the symbol grounding problem for words for which no direct experience with their referents is available. More specifically, we assumed that such words are grounded indirectly, via other, already grounded concepts with which they co-occurred or which they replaced in purely linguistic contexts. We expected that novel words presented in the context of existing words that elicit action-congruency effects would themselves elicit these congruency effects, by serving as a cue to re-activate these existing words, which then in turn re-activate sensorimotor traces.

We tested this assumption in four experiments implementing four purely linguistic learning phases: We presented novel words paired with different existing words from the same category (Experiment 1), we replaced existing words from the same category with novel words in natural texts (Experiment 2), we presented novel words as “translations” for existing words (Experiment 3), and we presented novel words as labels for new, combined concepts based on these existing words (Experiment 4). On the concept level, the novel words were therefore presented as referring to existing entities (Experiment 3), to novel entities created from combing two existing entities (Experiment 2 and 4) and to entities that were not further specified, or underdetermined (Experiment 1 and, to some degree, Experiment 2).

In all four experiments, we observed the same pattern of results: Contrary to our initial assumptions, we did not observe action-congruency effects for novel words that were learned in purely linguistic contexts (up- or down-contexts) in any of the four experiments. On the other hand, in all four experiments participants were able to correctly categorize the novel words as up- or down-words when they were forced to make an explicit judgement. This information necessarily had to be induced from the learning contexts the novel words were presented in, as this was the only available source of information on these novel words.

In principle, the fact that all four studies reported here yielded null effects poses a problem for their interpretability. However, we assume that the statistical power of our
studies was reasonably high, based on our initial power analysis based on the data by Lachmair et al. (2011) and Öttl et al. (2017). Therefore, we base the following discussion on the assumption that our studies show the absence of an action-congruency effect. This discussion will consist of two parts: We will first address methodological and technical issues, before we turn to a discussion of the implications of these results given they correctly indicate the absence of a congruency effect.

**Methodological Discussion**

First of all, participants did not necessarily have to read the words presented on the screen to react according to the words’ ink colour in the test phase task. However, as already shown in the study by Stroop (1935), reading is an automatic process that occurs even if task does not require it. Furthermore, our test phase task was identical to the one employed by previous studies that found action-congruency effects (Lachmair et al., 2011), also for novel words (Öttl et al., 2017). We therefore consider it unlikely that participants did not read the novel words in our test phases.

It could also be argued that participants did not learn anything at all in any of the learning phases, which would necessarily lead to an absence of congruency effects. However, the fact that, in all four experiments, participants were clearly capable of making correct judgements in the explicit judgement task renders this argument invalid. Another option is that participants did not store what they learned, or rapidly forgot it after the learning phase was finished. However, participants made their correct explicit judgements after the test phase was completed, which renders this option implausible as well.

Another issue might be that our learning phases were too short, or did not encourage deep processing, and therefore the novel words were learned too superficially to elicit a congruency effect. For example, participants never had to use the novel words for communicative purposes, as is the case for actual L2 words (Dudschig et al., 2014). However, the learning phases in our experiments took between 20 and 30 minutes to
complete and were not shorter than the learning phase employed by Öttl et al. (2017), albeit qualitatively very different. However, our learning phases were also not shorter than those in similar studies that found learning effects on variables such as categorization performances and similarity judgements (Ouyang et al., 2017; Lazaridou et al., 2017). This, in combination with the fact that participants were able to make correct judgements in the explicit judgement task, leads us to the assumption that the learning phases we employed were long enough to form a representation for the novel words that is clear enough to perform a variety of different tasks (Ouyang et al., 2017; Lazaridou et al., 2017). In fact, in none of the learning phases did the participants get any explicit information about the vertical location of the novel words they learned, but only encountered them in linguistic contexts implying a certain vertical location. Therefore, participants did not merely have to recall or retrieve information presented to them, but to induce the relevant information for the explicit judgements from the respective learning contexts. This strongly suggests that participants have indeed acquired conceptual knowledge about the novel words, which contradicts the assumption that participants learned the novel words too superficially.

Of course, more fundamental problems with our experiments can still be brought forward. One possibility is that the learning phases lead to conceptual learning, but were not designed adequately to lead to a re-activation of experiential traces for the relevant dimension in the test phases. Another, related possibility is that participants only re-activate the knowledge obtained in the learning phase when they are explicitly forced to do so, but not automatically, when this knowledge is task-irrelevant. Both possibilities imply that all the learning phases we employed induce some learning, but not the kind of learning needed to elicit automatic re-activations of experiential traces in the test phase. However, this does not exclude that other purely linguistic learning scenarios might lead to such an adequate kind of learning.

In fact, one can argue that our experiments and the study by Öttl et al. (2017) are two ends of a scale, with our experiments implementing learning from more or less shallow
linguistic experience, and the study by Öttl et al. implementing learning from direct
sensorimotor experience with novel word’s referents on the relevant dimension (vertical
location). Between these extremes, a large variety of other learning scenarios is possible.
For example, one could provide extensive and detailed linguistic explanations on the
meaning of the novel words, which could include mentioning a number of perceptual
features attributed to them. This could even include explicitly mentioning the vertical
location associated to the concepts described to the novel words. Another purely linguistic
learning scenario would be to include the novel words as important concepts in longer texts
or stories, for which simulation-based processing should play a more important role
(Barsalou et al., 2008).

On the other hand, one can provide participants with sensorimotor experience with
the novel word’s referents. If sensorimotor experience with some features of a novel word’s
referent is available (directly or indirectly, for example from pictures), it might be easier to
make inferences about other perceptual features from linguistic experience (Andrews,
Vigliocco, & Vinson, 2009). This could in turn lead to a stronger association with (and a
facilitation in the activation of) such experiential traces. Learning scenarios implementing
this approach can include presenting pictures or movies showing the word’s referents and
their perceptual features, or actually showing the actual referents, without presenting them
in the actual vertical location. Of course, one can also provide participants with both
linguistic and sensorimotor experience in a single learning scenario, such as showing the
referents and telling participants about their typical vertical orientation.

Under which of these learning conditions action-congruency effects are elicited by
novel words, if at all, remains an open and, above all, empirical question. On the basis of
the reasoning presented in the Introduction of this article, we strongly suspect this to be
possible in principle: After all, we seriously doubt that many participants ever made direct
experience with an actual submarine located in the lower vertical field, but the word
submarine clearly elicits action-congruency effects. Based on the results of our experiments,
we assume that, if purely linguistic information experience can produce these effects, this should only be possible in learning scenarios encouraging deep simulation-based processing of the novel word’s meanings, and not short-term learning scenarios as used in the current study. This might allow for direct associations to be formed between the novel words and sensorimotor experience re-activated for the simulation.

**Theoretical Discussion**

Up to this point, our discussion was implicitly based on our initial hypothesis, the assumption that novel words can in principle produce congruency effects after being learned from purely linguistic context. However, as counter-arguments can be brought forward against most methodological objections, there is also the possibility to draw a conceptual conclusion from the results we obtained: We did not find an action-congruency effect in any of our four experiments because our initial hypothesis does not hold, and learning novel words from purely linguistic context is not sufficient to elicit an action-congruency effect. The following Theoretical Discussion will now be based on this interpretation of our experimental results.

Following this interpretation, direct experience appears to be a sufficient (Öttl et al., 2017), but also a necessary condition for automatic activations of sensorimotor experiential traces upon encountering a word. This might be either because such traces are not activated upon encountering a word for which no direct experience is available, or because the word was never linked to experiential traces in the first place. However, if we assume that automatic activations of relevant experiential traces can indeed not be inherited via linguistic context, we are left with the initial problem: How can we understand words if we have no direct sensorimotor experience with their referents, and how does the embodied cognition account scale up to the entirety of our lexicon?

A possible solution to this problem is the conclusion that a re-activation of experiential traces is not actually necessary to understand words or concepts (contrary to
assumptions made by strong embodiment accounts, see Glenberg & Robertson, 2000; Glenberg, 2015a). This conclusion arises from the strong embodiment assumption that understanding concepts necessarily implies a re-activation of experiential traces. Conversely, without an activation of experiential traces there is no understanding of concepts. In terms of the study by Lachmair et al. (2011), correctly understanding the aspect of vertical orientation would therefore not be possible without an activation of the respective sensorimotor traces, which are observable in action-congruency effects. However, given our experimental results, it is difficult to claim that participants did not understand the meaning of the novel words, since they were perfectly able to make correct explicit judgements. Importantly, participants did not just understand some aspect of the novel concepts, but exactly the vertical dimension which was relevant here. Hence, a strong embodiment view does not seem to be compatible with our results, and a re-activation of experiential traces is apparently not necessary to understand every aspect of concepts.

This interpretation has consequences for theories of embodied language comprehension, such as the LASS model (Barsalou et al., 2008), which postulates an earlier-peaking linguistic processing of language stimuli that is followed by activation of embodied simulations such as the reactivation of sensorimotor experiential traces. In our experiments, we do not find evidence for this second step of processing, which should have been observable in action-congruency effects. Taking our results seriously, the conclusion can be drawn that this second step did not occur, with the implication that the activation of embodied simulations is not a necessary step for the processing of every word in every situation. According to Barsalou et al. (2008), both linguistic processing and embodied simulations always take place. However, these authors also argue that the relative role played by these two processes varies greatly depending on the task: In tasks that require only shallow processing, such as lexical decision tasks, linguistic processing of the stimuli is mostly sufficient, and therefore does most of the work (additionally, Connell & Lynott, 2013, provide evidence that linguistic information is indeed relevant also for deep
conceptual processing). However, the absence of action-congruency effects in our experiments cannot simply be attributed to task-dependent shallow processing of the stimuli, since we used exactly the same tasks as earlier studies in which action-congruency effects were observed (Lachmair et al., 2011; Öttl et al., 2017). Therefore, the absence of action-congruency effects in our experiments has to be attributed to the purely linguistic learning phases being not sufficient for resulting in these transfer effects of experiential associations. Nevertheless, these learning phases were sufficient for transferring non-experiential meaning aspects, which are clearly available in the explicit testing phase. This supports the interpretation that experiential associations are not activated in the processing and comprehension of all words.

In fact, this conclusion is in accordance with results from a line of research investigating whether the re-activation of experiential traces is functionally relevant for understanding, or merely a by-product of language processing (Fischer & Zwaan, 2008). Very recent results showing that language comprehension is not impaired when sensorimotor systems are occupied by other tasks call this functional relevance into question (Strozyk, Leuthold, Miller, & Kaup, submitted; Strozyk, Dudschig, & Kaup, 2017). Note however that the absence of functional relevance does not claim that concepts are not grounded in sensorimotor experience, but only that re-activating this experience is not necessary to understand concepts. This could also apply to the novel words in our experiments: It might still be possible that they are in some way indirectly grounded to experience via the linguistic context they occurred in, but if that is the case, this experience is not re-activated in the experimental settings we employed.

However, this leaves us with the question how participants were able to understand the aspect of vertical dimension for the novel words and to make correct judgements about it, if the answer that the vertical dimension is understood by re-activating the relevant experiential traces falls short. This points in the direction that a strong embodiment account might not be able to account for all phenomena associated to concepts, and that
amodal representations also play their part. In our experiments, participants could have just learned or inferred amodal, abstract propositions such as \textsc{location(up,mende)} in the learning phases, which they then retrieved for their explicit judgements. Under these conditions, one would not expect to observe any action-congruency effects in the test phases. Assuming (at least partially) amodal representations can also give efficient solution to the problem of scalability mentioned earlier, that our concept knowledge vastly exceeds the experience and information available to us (see Landauer & Dumais, 1997). One example for such a solution are distributional models that computationally derive word meanings from word co-occurrences in purely linguistic contexts (Landauer & Dumais, 1997; Lund & Burgess, 1996; Griffiths, Steyvers, & Tenenbaum, 2007).

In the literature on embodied cognition, there has been and still is a debate on how to account for concepts whose referents are not easily available for direct experience. This debate is mainly focussed on abstract concepts, such as \textit{freedom}, \textit{distance} or \textit{difference}, whose referents are not concrete objects that can be perceived and interacted with (e.g. Dove, 2016; Lakoff & Johnson, 1980; Louwerse, 2011; Thill, Padó, & Ziemke, 2014; Takano & Utsumi, 2016; Wiemer-Hastings & Xu, 2005; for a recent review, see Borghi et al., 2017). In this line of research, several approaches have highlighted the intermediating role played by language in the processing and representation of abstract concepts (Borghi et al., 2017 refer to these theories as \textit{multiple representation theories}). This is in line with the initial hypothesis of our study, where we also conceptualized language and linguistic experience as a cue to activate relevant sensorimotor experiential traces. In this context, we already discussed the LASS model (Barsalou et al., 2008), according to which an earlier linguistic processing of linguistic stimuli is followed by an activation of associated embodied simulations, with the relative role played by these two processes depending on the task to be performed. According to the LASS model, the processing and understanding of both concrete \textit{and} abstract concepts always involves an activation of embodied simulations. On the other hand, according to the representational pluralism view proposed by Dove (2011,
2014), both modal and amodal aspects play a role for the representation of abstract concepts. This view states that abstract concepts activate more disembodied linguistic information and associations than concrete ones, and therefore allows for concepts to be represented, at least partly, via this linguistic information (as theories of amodal, abstract symbols would assume). The view that abstract concepts activate more linguistic (and social) information and experience is also shared by the words as social tools (WAT) theory (Borghi & Binkofski, 2014), which also states that re-activating sensorimotor experiential traces is not sufficient to understand all concepts, especially not abstract concepts. However, in contrast to Dove’s representational pluralism and in line with the LASS model, the WAT theory assumes that sensorimotor experience has to be activated to understand meaning. A similar line of thought is implemented in the grounding and sign tracking approach (Prinz, 2002).

Hence, multiple accounts on the representation and processing of abstract concepts have been proposed, and postulated that language and linguistic information plays an important role for these concepts. However, we want to emphasize that the problem of unavailable referents, or of referents with which no direct experience is made, not only affects abstract concepts. In fact, we have the same situation also for a considerable number of concrete objects - concepts such as Carthage, bacteria, Charlemagne, stem cells or earth’s core. Therefore, theories of embodied cognition have to go beyond a simple concrete-abstract distinction in order to address the symbol grounding problem for the entirety of the lexicon.

Furthermore, our experiments show that the assumption of indirectly grounding concepts via language (see Barsalou et al., 2008; Louwurese, 2011; Thill et al., 2014; Takano & Utsumi, 2016, for proposals in this direction) is not as straightforward as one might initially suspect. We did not find evidence that any experience is re-activated when words learned from a purely linguistic context are encountered, and therefore no evidence favouring an embodied approach over a more classical, amodal approach, or over a
representational pluralism view. While this does not imply that there are no indirectly grounded words in principal, it suggests that not all the words we understand necessarily need to be grounded in experience, directly or indirectly.

To conclude our Theoretical Discussion, it is worth to once again come back to the initial proposals made by Harnad (1990) to address the symbol grounding problem. In this approach, Harnad explicitly stated that not all symbols necessarily have to be grounded in experience. Instead, a set of elementary grounded symbols is sufficient. On basis of this, Harnad then proposed a hybrid symbolic/nonsymbolic system, where the representation of some symbols incorporates nonsymbolic (or modal) sensorimotor experience, but other concepts are represented in a symbolic (or amodal) fashion, via their relations to other symbols. As stated above, hybrid models incorporating a representational pluralism have been proposed to address shortcomings of the embodied cognition approach (Dove, 2009, 2011, 2014), and in computational linguistics and cognitive science, models have recently been developed that integrate perceptual as well as linguistic experience in order to derive word meaning representations (Andrews et al., 2009; Bruni, Tran, & Baroni, 2014; Johns & Jones, 2012; Silberer, 2017). In hybrid models, the representation format of a word can be assumed to depend on the experience available with that word: If direct experience with the referent is available, the representation format is modal, and if the word has been learned only from its linguistic context, the representation format is amodal instead\(^9\). We propose that our results are compatible with such a hybrid view, with words meanings being represented in an amodal format if no direct experience with their referents is available.

\(^9\)Of course, one can also assume that each representation is partly modal and partly amodal, depending on the kind of experience that has been made
References


Dudschig, C., de la Vega, I., & Kaup, B. (2014). Embodiment and second-language:


Chicago Press.


Table 1

*Word material used for Experiment 1*

<table>
<thead>
<tr>
<th>Novel words</th>
<th>Up-ratings (prop.)</th>
<th>Up-words</th>
<th>Down-words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Essede</td>
<td>.50</td>
<td>Planet <em>(planet)</em></td>
<td>Keller <em>(basement)</em></td>
</tr>
<tr>
<td>Fente</td>
<td>.51</td>
<td>Dach <em>(roof)</em></td>
<td>U-Boot <em>(submarine)</em></td>
</tr>
<tr>
<td>Greites</td>
<td>.48</td>
<td>Wolke <em>(cloud)</em></td>
<td>Teppich <em>(carpet)</em></td>
</tr>
<tr>
<td>Emahle</td>
<td>.52</td>
<td>Hochsitz <em>(high seat)</em></td>
<td>Sumpf <em>(swamp)</em></td>
</tr>
<tr>
<td>Riehrer</td>
<td>.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seige</td>
<td>.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mende</td>
<td>.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gehaff</td>
<td>.47</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2

Distribution of novel words over learning contexts, by experiment. Numbers indicate for how many participants the respective novel words were assigned to an Up-context. The number of participants for which the respective words were assigned to a Down-context is given by \( n(\text{Down-context}) = 45 - n(\text{Up-context}) \)

<table>
<thead>
<tr>
<th>Novel word</th>
<th>Exp 1</th>
<th>Exp 2</th>
<th>Exp 3</th>
<th>Exp 4</th>
<th>Exp 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Essede</td>
<td>24</td>
<td>22</td>
<td>23</td>
<td>26</td>
<td>25</td>
</tr>
<tr>
<td>Fente</td>
<td>24</td>
<td>23</td>
<td>20</td>
<td>22</td>
<td>25</td>
</tr>
<tr>
<td>Greites</td>
<td>26</td>
<td>18</td>
<td>24</td>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>Emahte</td>
<td>19</td>
<td>17</td>
<td>18</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>Riehrer</td>
<td>22</td>
<td>20</td>
<td>25</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>Seige</td>
<td>20</td>
<td>25</td>
<td>22</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>Mende</td>
<td>21</td>
<td>27</td>
<td>22</td>
<td>27</td>
<td>26</td>
</tr>
<tr>
<td>Gehaff</td>
<td>24</td>
<td>28</td>
<td>26</td>
<td>16</td>
<td>18</td>
</tr>
</tbody>
</table>
Table 3

*Chimera pairs and example sentences for one chimera for Experiment 2*

<table>
<thead>
<tr>
<th>Learning Context</th>
<th>Chimera</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Up-Words</strong></td>
<td></td>
</tr>
<tr>
<td>Dach (<em>roof</em>)</td>
<td>+ Spitze (<em>peak</em>)</td>
</tr>
<tr>
<td>Hochsitz (<em>high seat</em>)</td>
<td>+ Dachbalken (<em>roof beam</em>)</td>
</tr>
<tr>
<td>Planet (<em>planet</em>)</td>
<td>+ Komet (<em>comet</em>)</td>
</tr>
<tr>
<td>Wolke (<em>cloud</em>)</td>
<td>+ Mond (<em>moon</em>)</td>
</tr>
<tr>
<td><strong>Down-Words</strong></td>
<td></td>
</tr>
<tr>
<td>Sumpf (<em>swamp</em>)</td>
<td>+ Pfütze (<em>puddle</em>)</td>
</tr>
<tr>
<td>U-Boot (<em>submarine</em>)</td>
<td>+ Taucher (<em>diver</em>)</td>
</tr>
<tr>
<td>Keller (<em>basement</em>)</td>
<td>+ Graben (<em>ditch</em>)</td>
</tr>
<tr>
<td>Teppich (<em>carpet</em>)</td>
<td>+ Gras (<em>grass</em>)</td>
</tr>
</tbody>
</table>

Examples

Es ist wieder einmal keine Seige (Wolke) am Himmel zu sehen.  

*Again, there is no Seige (cloud) to be seen in the sky.*

Es ist unklar, ob die Erde früher mehr als eine Seige (einen Mond) hatte.

*It’s unclear whether the earth formerly had more than one Seige (moon).*
Table 4

Phrases selected for the item material in Experiment 4, with mean rating scores on the scales Familiarity (1 - unfamiliar, 7 - familiar), Plausibility (1 - implausible, 7 - plausible), and Location (1 - lower, 7 - upper).

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Familiarity</th>
<th>Plausibility</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teppich aus Jeans (carpet made of jeans)</td>
<td>1.63</td>
<td>5.87</td>
<td>1.23</td>
</tr>
<tr>
<td>U-Boot mit Kapelle (submarine with a chapel)</td>
<td>1.73</td>
<td>4.43</td>
<td>1.47</td>
</tr>
<tr>
<td>Sumpf im Bergwerk (swamp in a mine)</td>
<td>2.13</td>
<td>4.37</td>
<td>1.53</td>
</tr>
<tr>
<td>Keller mit Guckrohr (basement with a periscope)</td>
<td>2.33</td>
<td>4.80</td>
<td>1.80</td>
</tr>
<tr>
<td>Dach aus Seide (roof made of silk)</td>
<td>2.10</td>
<td>4.30</td>
<td>5.53</td>
</tr>
<tr>
<td>Planet aus Kalk (planet made of chalk)</td>
<td>2.17</td>
<td>4.83</td>
<td>6.07</td>
</tr>
<tr>
<td>Hochsitz auf Turm (high seat on a tower)</td>
<td>2.23</td>
<td>3.67</td>
<td>6.30</td>
</tr>
<tr>
<td>Wolke auf Mars (cloud on Mars)</td>
<td>2.07</td>
<td>3.60</td>
<td>6.53</td>
</tr>
</tbody>
</table>
Figure 1. Left panel: The experiential trace model by Zwaan and Madden (2005). Lines indicate learned associations between perceptual experience (left) and linguistic experience (right), and within these categories. Right panel: Our proposal for a version of the experiential trace model, with no perceptual experience available associated to the word “Mende”. “In this case, “Mende” is indirectly grounded via “Ocean”. The dashed line indicates a direct association that might emerge.
Figure 2. Mean Reaction times by word type and response direction in the pre-test study. The filled circles on the left of each bar show the distribution of mean reaction times by participants, the empty circles on the right of each bar show the distribution of mean reaction times by item.
Figure 3. The experimental setup as seen from a participant’s perspective. The overlay’s buttons are placed over the \textit{TAB}, \textit{u}, \textit{o} and \textit{END} keys of the keyboard.
**Figure 4.** Left panel: Mean Reaction times by learning context and response direction in Experiment 1. The filled circles on the left of each bar show the distribution of mean reaction times by participants, the empty circles on the right of each bar show the distribution of mean reaction times by item. Right panel: Proportion of upwards-location ratings for novel words by learning context in Experiment 1, with .95 confidence intervals.
Figure 5. **Left panel:** Mean Reaction times by learning context and response direction in Experiment 2. The filled circles on the left of each bar show the distribution of mean reaction times by participants, the empty circles on the right of each bar show the distribution of mean reaction times by item. **Right panel:** Proportion of upwards-location ratings for novel words by learning context in Experiment 2, with .95 confidence intervals.
Figure 6. **Left panel:** Mean Reaction times by learning context and response direction in Experiment 3. The filled circles on the left of each bar show the distribution of mean reaction times by participants, the empty circles on the right of each bar show the distribution of mean reaction times by item. **Right panel:** Proportion of upwards-location ratings for novel words by learning context in Experiment 3, with .95 confidence intervals.
Figure 7. 

**Left panel:** Mean Reaction times with standard errors by learning context and response direction in Experiment 4. **Right panel:** Proportion of upwards-location ratings for novel words by learning context in Experiment 4. Error bars indicate standard errors for the proportion estimates.
Figure 8. **Left panel:** Mean Reaction times with standard errors by learning context and response direction in Experiment 5. **Right panel:** Proportion of upwards-location ratings for novel words by learning context in Experiment 5. Error bars indicate standard errors for the proportion estimates.