Author:
Ayravainen, Laura

Title:
Learning Spoken Words from Context

Learning Spoken Words from Context – Effects of First and Second Language Proficiency on Word Form and Word Meaning Acquisition

Laura Eeva Maria Äyräväinen

School of Psychological Science

University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Master of Science by Research in the Faculty of Life Sciences

September 2018

Word count: 24,511

Abstract

Acquiring new vocabulary in adulthood happens largely from context. The efficiency of contextual learning and the quality of lexical representations formed through it has been explored in the visual domain (Frishkoff, Perfetti & Collins-Thompson, 2010). However, contextual learning in the auditory domain and the development of word form and word meaning representations in such settings has received little attention. Additionally, high language proficiency facilitates novel word acquisition (Shefelbine, 1990; Kaushanskaya, 2012), but the type of word learning facilitated and the source of this effect are still unclear. To address these gaps in research, word learning was investigated in L2 learners with aurally presented sentences where novel word meanings were inferred from the context. Whether this learning task leads to novel word representations that are integrated into the mental lexicon (i.e. lexicalized) was tested immediately and 48 hours after learning. Explicit and implicit novel word knowledge was examined with two-alternative forced-choice, Pause Detection and Semantic Relatedness task, using behavioral and ERP measures of learning. In the Semantic Relatedness task, lexicalization of the novel words was tested via semantic priming: if novel words have been lexicalized, they should prime their meanings (e.g. *cathedruke* - *basket*) as well as their semantic associates (e.g. *cathedruke* - *weave*). Our findings demonstrate above-chance recognition accuracy for novel word forms and their meanings immediately after learning; moreover, recognition accuracy increased between the testing sessions and varied as a function of L1 proficiency. Higher L2 proficiency was associated to higher gains in word form recognition accuracy over time. Importantly, semantic priming (indexed by N400) was found for both meaning and semantic associate targets 48 hours after learning. The findings contribute to the discussion of language proficiency effects in word learning and of how lexical representations develop from context.

Word count: 290

Dedication and Acknowledgements

I am deeply grateful to Aran Barbal Staab for his support on every level throughout this project. He was always able to accommodate what was best for me.

I also wish to express my immense gratitude to Howard Staab for his generous financial support during this project.

Author’s declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE:.....

Table of Contents

List of Tables and Figures

1. Introduction

We learn new words throughout our lives and a large part of our acquired vocabularies are words we have learnt from context. (Sternberg, 1987). Repeated exposure to a word in varied contexts allows us to form a stable representation of it, from partial word knowledge to a fully lexicalized item. A general meaning of a word can be understood, especially in supportive contexts, even when defining the word might still be impossible. This is seen in how people understand passively more words than they can actively use (Durso, & Shore, 1991). However, ability to explicitly recognize or recall a word seems to not be the end of the story. What is special about word learning, is that newly acquired concepts and the labels for them are eventually integrated into the network of concepts and labels we already know, i.e. our mental lexicon (Davis & Gaskell, 2009). As a result, the newly learnt items relate to the existing items in a way that in part defines how we conceptualize the world around us. Words that are frequently used together tend to form stronger connections, thus reflecting the way our environment is organized as well as how our language is organized. (Hagoort, Hald, Bastiaansen, & Petersson, 2004).

Understanding the process of a newly learnt word becoming a fully functioning part of the mental lexicon is important for several reasons. The most immediate implication is that understanding the process of word learning allows the development of more efficient educational practices for language learning. On a more general level, understanding how language is processed in healthy population informs the development of diagnostic tools and interventions for pathological language processing. An even broader implication of elucidating the processes of novel word acquisition is a better understanding of learning in general, from learning strategies to the nature of memory systems.

The current study aims to explore the nature of lexical representations of novel spoken words that have been learnt from context. Additionally, the effects of language proficiency in contextual novel word acquisition are investigated. Relevant literature is reviewed first,

followed by an overview of the current study with a more detailed description of the aspects of the study that are adding to the body of knowledge in this area of research.

1.1 Novel word acquisition

Understanding novel word acquisition requires a framework for the broader phenomena it is a part of. Known words are stored in the long-term memory and as such, the way they are accessed and the way they interact with other items in memory is an important part of the bigger picture in investigations of novel word acquisition, i.e. adding new items to the network of already known words. For the purposes of the current study, the theory of Spreading Activation of Semantic Processing (Collins & Loftus, 1975) serves as a more general account for how word knowledge is organized. According to this theory, items in memory are represented as nodes that have different properties. The nodes with more shared properties (e.g. members of the same semantic category) have more connections between each other. Activation of one node results in activation of other nodes that are connected to it and the degree to which the connected nodes are activated depends on the strength of connections between the two nodes. Regarding known words, this kind of connectivity applies to phonetic and orthographic as well as semantic properties of the words: the word forms, i.e. the names of the concepts, and the word meanings i.e. the concepts, are organized according to similarity. There are two important effects in language processing that are considered instances of spreading activation within the network of word forms and word meanings (i.e. the mental lexicon). At the level of word meanings, one such effect is the semantic priming effect (e.g. Meyer & Schvaneveldt, 1971; Fischler, 1977; Lucas, 2000), where previously presented item (e.g. a word) facilitates the subsequent processing of a related item. For example, a word is responded to quicker when preceded by a semantically related word (e.g. *dog* followed by *cat*) than when preceded by an unrelated word (e.g. *dog* followed by *car*). [1] At the level of word forms, the co-activation of similar lexical items is believed to be a central element in spoken word recognition. This co-

[1] Semantic priming effect can arise from either association between the lexical items (e.g. *mouse* and *cheese*) or from purely semantic similarities between items (e.g. *mouse* and *shrew*). These are not separated for the purposes of brevity in the current example.

activation is referred to as lexical competition: when a speech signal unfolds in time, word forms that match it co-activate and compete for activation until the best match for a given speech signal is selected, i.e. a word is recognized. One model for lexical competition is the Cohort Model (Marslen-Wilson, 1987), which proposes that encountering the initial segments of a speech signal activates all the word forms in the mental lexicon with a matching onset to these segments (the initial cohort set). As the speech signal progresses, the candidate word forms are gradually eliminated and a word is recognized when only one word form matches the given speech signal. For words presented in isolation, their recognition happens at the uniqueness point of a word, the point at which the onset of a word form in the mental lexicon no longer matches that of any other word form (e.g. the uniqueness point of a word *captive* is at the final phoneme /v/ as it diverges from its cohort *captain* at this point). According to the cohort model, the recognition speed of a spoken word depends on its uniqueness point and the size of the initial cohort set, with early uniqueness points and small initial cohort sets leading to quicker recognition. The processing speed of lexical items can thus be affected by other items in the mental lexicon, as exemplified by the semantic priming effect and the lexical competition. With this type of organization and interaction of well-established items in mind, adding new items to the mental lexicon will be considered next.

The process of learning new words happens gradually and there are different aspects of word knowledge that emerge at different stages of this process (Lindsay & Gaskell, 2010). The Complementary Systems Account of word learning (Davis & Gaskell, 2009) identifies two stages of novel word acquisition: an initial familiarization stage, supported by medial temporal and hippocampal learning, and a lexical consolidation stage, where new words become cortical representations. The first stage is akin to forming an episodic memory trace, which is sufficient for explicit recognition and recall of novel lexical items while these items remain separate from long-term memory, whereas the second stage is characterized by an integration of lexical items into the existing phonological and semantic network in a way that allows interaction between the new and older items to take place. Such interaction was described above as the semantic priming effect and engagement in lexical competition. Thus, a word can be considered a fully functioning member of the mental lexicon (i.e. lexicalized) when it exhibits behavior

characteristic to well-established, known words: when it can affect the processing of other known words. The complementary systems account of word learning also suggests that sleep plays an important role in memory consolidation of novel lexical items from episodic memory traces to longer-term word knowledge. Investigations of word learning in the current study are approached from the perspective of the complementary systems account of word learning.

As integration of novel words into the existing mental lexicon is a crucial part of word learning, recent studies have used measures of lexicalization to investigate novel word acquisition in adults. Utilizing explicit measures of word knowledge, such as recognition and recall accuracy of novel words, and the above mentioned measures of lexicalization (i.e. the semantic priming effect and engagement in lexical competition), these studies have revealed several important aspects of novel word acquisition. They also provide evidence in favor of the complementary systems account of word learning. In a series of experiments, Gaskell & Dumay (2003) investigated lexicalization of newly learnt words by looking at the effects these new items might have on recognition of known words. Participants were exposed to novel word forms that resembled existing word forms (e.g. *cathedruke* as a novel word based on an existing word *cathedral*). The existing word forms had an early uniqueness point and the novel words differed from the existing words only at the end. Lexicalization of the novel words would thus move the uniqueness point of the existing words further towards the end. This would slow down the recognition of the existing words (critical words) compared to control words (existing words matched with critical words in length, frequency of occurrence and uniqueness point but for which no novel word forms that resembled them had been presented). A novel words' ability to engage in lexical competition, i.e. to slow down recognition of known words that it resembles, will be referred to as the lexical competition effect. In Experiment 1, lexical decision task (LDT, speeded decisions on whether a presented item is a word or not) and two-alternatives forced-choice task (2AFC, choosing the correct novel word from two highly similar options) were used to investigate the immediate effects of learning (a distraction task was administered between the learning phase and the learning tests). Participants' explicit recognition accuracy of the novel word forms was high (over 90% on average) in the 2AFC task, whereas the reaction times (RTs) in the LDT showed faster responses to the critical words compared to control words, an

opposite pattern to what was expected had the novel words been lexicalized. The time course of lexicalization of novel words was investigated in Experiment 2. Participants were exposed to novel word forms on 5 consecutive days; on each day learning was assessed using the same tests as in Experiment 1. The explicit recognition of the novel words in the 2AFC task was again high both immediately and throughout the 5 testing sessions. In the LDT, the RTs were found to be reliably slower on day 4 and 5 for critical words compared to control words, thus showing the lexical competition effect, suggestive of lexicalization of the novel word forms. Finally, in Experiment 3, participants were exposed to the novel word forms and tested immediately and then a week later with no further exposure in between the two testing sessions. The amount of exposure to the novel word forms in the first testing session was equivalent to the combined amount of exposure during the first 3 sessions in Experiment 2 (e.g. each novel word was heard 36 times). In Experiment 3, the explicit recognition accuracy of the novel words was tested with 2AFC task, but the emergence of the lexical competition effect was investigated utilizing a Pause Detection task (PD; Mattys & Clark, 2002). In the PD task, participants make speeded decisions on whether a presented word has a pause in it or not. Mattys & Clark (2002) found longer RTs to words with late uniqueness point and non-words with high number of initial cohorts, thus suggesting that this task is sensitive to the on-going lexical competition at the time the pause occurs. The advantage of this task is that performing well only depends on analysis at the level of phonemes, whereas the LDT might result in semantic processing of the stimulus words. As such, the influence of explicit word knowledge should be minimal in the PD, but not necessarily in the LDT. Taking these considerations into account, Gaskell & Dumay inserted 200 ms pauses in the existing words (critical and control words) and expected RTs to the critical words to be longer if the novel words had been lexicalized. Their results showed that the responses to critical words were not different from responses to control words in the first testing session, but they were significantly slower 1 week later. The explicit recognition accuracy in the 2AFC task was again high (over 96% on average) in both testing sessions. This series of experiments demonstrates that the explicit recognition of novel words emerges immediately, whereas the novel words engage in lexical competition only after a delay. One further question from these experiments was the nature of this delay: whether lexicalization of

novel words requires overnight sleep or whether time passing in general is sufficient. The role of overnight consolidation in acquiring novel words was investigated by Dumay & Gaskell (2007). In their experiment, participants were exposed to novel word forms and tested immediately (Session 1), 12 hours later (Session 2) and 24 hours later (Session 3). Half of the participants had slept before the Session 2, whereas the other half had stayed awake during the 12 hours between the Session 1 and the Session 2. Both groups were tested again 12 hours later (Session 3), at which point all the participants had had at least one night's sleep. Explicit recognition in the 2AFC task was high in all testing sessions and both groups. However, the lexical competition effect in the PD task emerged gradually: it was not found in either of the groups in Session 1. In Session 2, it was only present in the group that had slept before this testing session and in Session 3, the effect was observed in both groups. This experiment thus demonstrates how overnight sleep can affect the lexicalization of novel words.

The experiments described above focused on the acquisition of novel word forms only. However, a more realistic simulation of novel word acquisition is a task that requires linking the new word form to a meaning. Studies utilizing such a task (Dumay, Gaskell & Feng, 2004; Henderson, Devine, Weighall & Gaskell, 2015) report the same pattern of results regarding the lexical competition effect as the studies where no meaning was attached to the novel word forms (Gaskell & Dumay, 2003), although the time course of the emergence of the lexicalization effect varies. In a study by Dumay, Gaskell and Feng (2004, Experiment 1), participants were exposed to novel word forms with or without meanings attached to them and their explicit and implicit knowledge of these words was tested immediately (Session 1), 24 hours (Session 2) and 1 week later (Session 3). In the first session, there was an exposure phase to the novel words, followed by tests of novel word learning with LDT, 2AFC task and a free association task. The same tests were used in Session 2, followed by another exposure phase. Finally, the same tests were administered in Session 3. The LDT allowed detection of both lexical competition effect (as a difference in RTs to critical and control words described above in Gaskell & Dumay, 2003) and semantic priming effect (as a difference in RTs to known words that were semantically related or unrelated to a preceding novel word). In line with previous findings, neither learning mode resulted in immediate lexicalization effects. However, the novel word forms learnt only

based on phonology showed the lexical competition effect in LDT 24 hours and a week after the learning phase, whereas this effect was observed in word forms learnt with their meanings only a week later. Interestingly, the semantic priming effect emerged simultaneously with the lexical competition effect: in LDT, the semantic priming effect was observed only a week later. Similarly, in the free association task, the probability for participants to produce the meaning of the novel words increased only a week later on the expense of probability for producing the meaning of the base word. These results suggest that more time is needed for lexicalization effects to emerge when novel words are learnt with a meaning than when the words are learnt as a word form only. However, Henderson, Devine, Weighall & Gaskell (2015) found the lexical competition effect for novel word forms with the PD task already 24 hours after the initial exposure to the novel words, when the words were learnt in a sentence context (i.e. when the meanings of the novel words could be inferred from the context). As an addition to these studies, which utilized the lexical competition effect as a measure of novel word consolidation, an investigation focusing on the consolidation of meaning of the novel words was carried out by van der Ven, Takashima, Segers & Verhoeven (2015). In their study, participants studied existing low-frequency words paired with their definitions for a maximum of 2 hours or until they felt they had learnt all the words. A four-alternatives forced-choice for word definitions (recognition test) and a primed LDT (semantic priming test) were administered immediately after the learning phase and then repeated 24 hours later. Both tasks were presented visually. The prime-target pairs in the LDT were a novel word form paired with a semantically related word or an unrelated word. Crucially, the related target words had never been presented as part of the definition of the novel words in the learning phase (this aspect of the experimental design will be discussed further in the Current Study section). A near ceiling performance for recognition accuracy was found in both testing sessions, whereas the semantic priming effect only emerged 24 hours after the learning phase.

The results from these studies support the predictions from the complementary systems account of word learning. Firstly, the word form lexicalization was demonstrated to take place whether words are learnt as isolated word forms or as word forms associated with meanings. Secondly, semantic priming effect was found after a delay, but not immediately following

exposure to novel words. Thirdly, contrary to the lexicalization effects for word form or word meaning, explicit recognition of word form and meaning can be highly successful immediately after exposure to the novel words.

1.2 Contextual novel word acquisition

Learning novel words from context is a natural way of increasing one's vocabulary both in first language (L1) and in second language (L2). Instead of formal instruction and dictionary definitions, many words are indeed learnt from context. Furthermore, word acquisition by encountering a word in varied contexts is suggested to provide richer semantic representations of the novel word (Beck, McKeown & Kucan, 2002). Many recent studies investigating contextual word learning focus on the semantic representations of the newly acquired words. In a number of these studies, apart from behavioral data, brain responses are recorded using electroencephalography (EEG) to provide another index of lexicalization of the novel words. Event-related potentials (ERP) are brain responses to a stimulus category of interest, e.g. newly learnt words, obtained by extracting time-locked responses to individual stimuli from the continuous EEG data. An ERP component called N400 has been used widely in language research to study semantic processing because of its sensitivity to semantic manipulations. N400 is a negative going deflection peaking roughly at 400 ms after stimulus onset with a latency range of approximately 200-600 ms[2] and centro-parietal scalp topography (Kutas & Federmeier, 2011). The N400 effect was found first as a stronger negative-going deflection to unexpected final words of a sentence (*He takes his coffe with sugar and dog*) compared to expected sentence endings (*He takes his coffee with sugar and milk*) by Kutas & Hillyard (1980). It has since been used in a variety of settings investigating semantic processing. The N400 effect is considered to reflect meaning congruence between an item and its preceding context. In the semantic priming paradigm, a target word preceded by a semantically related prime word elicits a smaller (i.e. less negative-going) N400 response than a target preceded by an unrelated word (e.g. Bentin, McCarthy & Wood, 1985; Holocomb, & Neville, 1990).

[2] Typically observed N400 latency is 250-550 ms in young adults (Kutas & Federmeier, 2009).

In a study by Frishkoff, Perfetti & Collins-Thompson (2010), novel word acquisition from written context was investigated utilizing behavioral and ERP (N400) indices of lexicalization. Participants were presented with sentences that each contained an unknown word (an existing low frequency word). The sentence contexts were either informative, where the context supported inferring the meaning of the novel word, or uninformative, where the context provided less support for the meaning inference. Recognition test where participants chose the correct meaning out of 5 options for each novel word was administered immediately after the learning phase and 48 hours after the learning session (Session 2). In Session 2, participants also completed a semantic relatedness task (SR, speeded decisions about whether presented two words are semantically related or unrelated) while their reaction times (RTs) and ERPs were recorded. The recognition tests showed successful recognition immediately and after a delay for words learnt in both informative and uninformative contexts, but the informative contexts produced higher recognition accuracy. Whereas the RT data did not show the expected semantic priming effect, the N400 was smaller in related vs unrelated pairs – both for known words and for novel words trained in informative contexts. A similar but weaker N400 effect was also found for novel words learnt from uninformative contexts. Although the N400 effect was not measured immediately after learning, detecting it after 48 hours along with successful explicit recognition performance demonstrate lexicalization of the novel words in line with the complementary systems account of word learning. In contrast to these findings, a surprisingly fast lexicalization effect was reported by Mestres-Misse, Rodriguez-Fornells & Munte (2007). In their series of experiments, participants were presented with written sentences with a novel word (a non-word, e.g. *lankey*) in each. The context for novel words was either informative or uninformative. Sentences that consisted of known words were also presented as a control condition. Each novel word was presented in 3 different contexts and ERPs were recorded simultaneously. The N400 response to the novel words in the informative contexts grew progressively smaller (less negative) during the presentation of the 3 sentence contexts, whereas no such reduction was found for responses to the novel words in uninformative contexts. The responses to learnt novel words (informative condition only) were indistinguishable from those to known words by the third encounter of the novel words.

Meaning acquisition was also tested with the SR task immediately after the learning phase. The SR task revealed longer responses to related word pairs for both known words and novel words (i.e. an opposite to semantic priming effect), but the N400 effect was found for both known and novel words learnt in informative contexts. This immediate N400 effect will be considered further in the Discussion.

These studies demonstrate that the emerging lexicalization of novel words cannot always be detected using behavioral measures, but more sensitive measures (such as ERPs) can show lexicalization effects earlier, in some cases immediately after learning. Whereas these ERP studies investigated contextual learning in the visual domain, a study by Henderson et al. (2015) mentioned in the previous section (Novel word acquisition, p. 7), demonstrated a lexicalization effect of novel word forms learnt from aurally presented sentences. This effect was found in Pause Detection task 24 hours after listening to the stories in which the novel words occurred and evidence of lexicalization of the novel words was found both in children and in adults.

Another line of research in contextual word learning comes from reading studies. Reliable novel word acquisition has been reported even in incidental learning conditions, where the purpose of the task was not revealed to be learning unknown words (Swanborn & De Glopper, 1999; Nagy, Herman, & Anderson, 1985). Additionally, print exposure has been found to be strongly associated with vocabulary size even when general ability measures (e.g. non-verbal reasoning skills) are controlled for (Stanovich & Cunningham, 1992). These studies support the claim that novel words can be learnt from context and such learning leads to lexicalization of the novel words, both in terms of word form and meaning.

1.3 Language proficiency in novel word acquisition

1.3.1 First language proficiency

The role of language proficiency in novel word acquisition is complex. Although proficient language use covers a variety of skills, vocabulary size is considered a good single measure of language proficiency (Staehr, 2008). Studies of reading and word learning from context provide evidence for positive effects of L1 proficiency, measured as vocabulary size or as reading ability.

A number of studies have shown how highly skilled readers learn new words from context more efficiently than less skilled readers (Jenkins, Stein & Wysocki, 1984; Cain, Oakhill & Bryant, 2004; Bolger, Balass, Landen & Perfetti, 2008). Reading ability itself is fundamentally dependent on vocabulary size and the quality of vocabulary knowledge (for L1: Tannenbaum, Torgesen & Wagner, 2006; For L2: Schmitt, Jiang & Grabe, 2011). This idea is expressed in the Reading Systems Framework (Perfetti & Stafura, 2014), according to which the mental lexicon is a central connection between word identification and comprehension, including later inferences from the text as a whole. While the relationship between reading ability and vocabulary size is difficult to disentangle, it is plausible to think of vocabulary as a prerequisite to reading ability. As such, although the studies mentioned above didn't explicitly report vocabulary size differences between more and less skilled readers, it can be conjectured that a large part of the word learning benefits associated to reading skills are mediated by large vocabulary and high quality lexical representations of the skilled readers. This issue is addressed more directly in studies of children listening to stories in kindergarten, where word learning benefit was found for children with larger vocabularies compared to their peers with smaller vocabularies (Sénéchal, Thomas & Monker, 1995; Ewers & Brownson, 1999). Additionally, Shefelbine (1990) reported word learning benefit for sixth graders with larger vocabularies over their peers with smaller vocabularies in contextual word acquisition task where students were intentionally trying to infer the meanings of the unknown words while listening to texts in which they occurred. Larger vocabulary was associated with higher recall and recognition immediately after the exposure to the novel words. Large vocabulary students learnt more words even considering that they had less unknown words to learn from the texts. Additionally, non-verbal reasoning skills did not predict performance in word learning. The last point is important in the light of two positions generally taken to explain the positive correlation between reading skills and vocabulary size: one suggests that large vocabulary facilitates reading ability, whereas the other proposes a common underlying factor that facilitates the development of both. Jensen's argument (as cited in Shefelbine, 1990) for higher reasoning skills in higher vocabulary students would thus not be supported by findings reported by Shefelbine.

Unlike studies of learning from context, Perfetti, Wlotko, & Hart (2005) investigated the effects of reading ability on flash card type word learning. Participants were divided into groups of more and less skilled readers based on their score on a reading comprehension task. Participants learnt 60 novel words (existing, low-frequency words) with their definitions in a 45 minute training session. Immediately after the training, the participants completed SR task, while their RTs and ERPs were recorded. Trained novel words, untrained novel words and untrained familiar words were presented in this task. Skilled readers outperformed less skilled readers in accuracy of semantic judgments for trained novel words, but not for familiar or untrained novel words. The RTs in this task were numerically faster for related words in familiar and trained novel word conditions, but no statistical significance was reported regarding this difference. Semantic priming as reflected by the N400 effect was found in the SR task for trained novel words and familiar words but not for untrained novel words. Furthermore, skilled readers showed a larger N400 effect than less skilled readers. These findings thus suggest an immediate lexicalization of novel words as measured by the N400 effect and at least trending behavioral semantic priming effect. Potential reasons for these immediate effects might be the extensive training of the novel words, in addition to perhaps motivating element of the learning task as the novel words were real words that could be useful outside of the laboratory. In any case, the study demonstrates an association between high reading skill and efficient novel word acquisition both in explicit recognition accuracy and electrophysiological responses.

The studies above have demonstrated a within-language proficiency benefit in word learning. That is, L1 proficiency facilitates learning new words in L1. In addition to this, Knight (1994) demonstrated effects of L1 proficiency in L2 word acquisition from context. In her study, L2 (Spanish) students were divided into high and low verbal ability groups (based on their American College Test scores in verbal ability, in L1 (English)). They read L2 texts without knowledge of the up-coming word learning tests. Tests of recall, where the participants provided L1 translation equivalents of the novel words, and recognition, where the participants chose the correct L1 definition out of 4 options, were administered immediately after reading the L2 texts and 1 week later. High verbal ability group outperformed the low verbal ability group in immediate and delayed recall and recognition of the novel words.

1.3.2 Second language proficiency

Several recent studies have reported that bilinguals outperform monolinguals in novel word learning tasks (Kaushanskaya & Marian, 2008, 2009; Kaushanskaya, 2012; Kan, Sadagopan, Janich & Andrade, 2014). Word learning in these studies was predominantly investigated with explicit recall and recognition tests where recall test involved producing the L1 translation equivalents of the novel words and recognition test involved choosing the L1 equivalent of the novel word out of 5 options (all participants had the same L1). Although immediate test performance in these studies was always higher in the bilingual groups, the same tests administered 1 week later have sometimes not shown difference between bilinguals and monolinguals (Kaushanskaya & Marian, 2008; Kaushanskaya, 2012; Kan & Sadagopan, 2014). While studies above tested highly proficient bilinguals who acquired their L2 early in life, second language proficiency or late bilingualism has also been shown to be advantageous for novel word acquisition (Papagno & Vallar, 1995; Van Hell & Mahn, 1997; Elgort, Perfetti, Rickles & Stafura, 2014; Nair, Biedermann & Nickels, 2016). Majority of these studies of late bilingualism also utilized explicit measures of memory and as in studies of early bilingualism, the method of learning novel words was direct instruction, where novel words were paired with a corresponding picture or a translation equivalent. Elgort et al. (2014) break this pattern with their study where more and less proficient L2 learners were taught novel L2 words in visually presented sentence contexts. The learning was tested with semantic relatedness task utilizing both behavioral and ERP measures. The testing session took place a day after the contextual learning phase. Participants saw some of the same sentences as in the learning phase and new sentences where the novel words were presented in either congruous or incongruous contexts. After the last word of the sentence was presented, a semantic probe appeared on the screen and participants had to judge whether this word was related to the final word of the just seen sentence. Response accuracies in this task were higher for the high proficiency group. The RTs revealed a semantic priming effect and the ERP data showed an N400 effect for both groups, but these lexicalization effects were reliably larger in the high proficiency group. Elgort and colleagues thus provide first evidence of more efficient lexicalization of novel L2 words that is associated with higher L2 language proficiency. Contrary to the above mentioned studies,

Bartolotti and Marian (2012) found no bilingual benefit in novel word acquisition task designed to elicit within-language interference. In their study, participants learnt non-words that overlapped with existing L1 words. During the learning phase, aurally presented novel words were paired with pictures (e.g a picture of an acorn paired with a novel word *shundo*, which overlaps with the English word *shovel*). The word learning test consisted of auditory presentation of the novel word form together with two pictures, one correct referent of the novel word and one competitor picture that referred to the L1 word that overlapped with the novel word form. There were no differences in accuracy or speed of picture – word form matching between monolingual and bilingual groups. However, the authors found that bilinguals did resolve cross-linguistic interference quicker than monolinguals, as shown by eye-tracking and mouse-tracking analyses, where monolinguals looked and directed the mouse towards the competitor picture more than bilinguals did.

With these findings, the underlying reason behind potential bilingual benefit in word learning is still unknown. More efficient novel word acquisition might be based on learning the new word forms more efficiently, or learning to link the new word forms to their meanings more efficiently. The former option seems plausible given that long-term phonological knowledge has been shown to influence phonological short-term memory (STM) performance in word learning (Gathercole, Frankish, Pickering & Peaker, 1999; Majerus, Van der Linden, Mulder, Meulemans & Peters, 2004; Majerus, Poncelet, Van der Linden & Weekes, 2008). Learning foreign word forms might be more demanding for monolinguals, whose phonotactic repertoires are solely L1-based and therefore smaller, than for bilinguals. Larger phonological network of a bilingual might provide more support for phonological STM when learning new phonological sequences. From this stand point it is not surprising that the majority of the bilingual word learning studies demonstrate a bilingual benefit in learning phonologically unfamiliar novel word forms, which essentially represents foreign vocabulary acquisition. There are, however, a few studies that show a bilingual benefit over monolinguals even in phonologically familiar novel word learning (Kan et al. 2014; Kaushanskaya, 2012; Kaushanskaya & Rechtzigel, 2012). These findings would suggest that bilingualism enhances both native and non-native word form acquisition through a mechanism which cannot be reduced to merely superior phonological knowledge or

phonological STM capacity. For instance, Kaushanskaya (2012) found in her study that bilinguals outperformed monolinguals in a native and non-native word learning task even when the two groups were matched in their phonological STM capacity, and suggested as one explanation that bilinguals are more efficient at linking new word forms to existing meanings or encoding multiple word forms to the same meaning – the kind of exercise bilinguals have more experience in due to second language acquisition. Kan et al. (2014) found that bilinguals outperformed monolinguals in novel word learning task where novel non-native or native word forms were paired with pictures of novel objects. This benefit was only found in comprehension measures of the novel word learning test (e.g. choosing the right novel object when the novel word was presented) and not in production (e.g. naming the presented novel object). Instead, when participants had to name novel objects in their native language and in a non-native language, the performance of bilinguals and monolinguals did not differ. This finding is particularly interesting given that two thirds of the monolinguals and an equal proportion of bilinguals had received speech practice training for the novel word forms in each language, where participants heard and repeated the native and non-native word forms. As the speech practice, which essentially familiarizes participants with the novel word forms, did not benefit either group more than the other, it seems that the difference in novel word learning between the groups did not lie solely in the efficiency of learning word forms. The bilingual benefit in learning foreign word forms has been demonstrated in several studies. The two studies described above suggest that the same applies to learning native word forms, although the pattern of results is not clear (Kaushanskaya, 2012; Kan et al., 2014; but see Kaushanskaya, Yoo & Van Hecke, 2013). If the bilingual benefit in novel word learning does not depend solely on more efficient learning of word forms, a closer look at the efficiency of linking the word form to a meaning is needed. Suggestions for more efficient semantic learning as the basis of bilingual word learning advantage have been made in two studies. Kaushanskaya and Rechtzigel (2012) tested the concreteness effect, i.e. the finding that concrete words are remembered better than abstract words, in novel word acquisition. Monolingual and bilingual participants learnt concrete or abstract meanings for novel word forms that followed L1 phonotactic rules (i.e. non-word – L1 translation equivalent pairs). A recall test was administered immediately after

the learning phase. The groups did not differ in recall accuracy for the abstract L1 translations of the novel words, but bilinguals recalled significantly more concrete L1 translations than monolinguals. Furthermore, while the concreteness effect was found in both groups, it was significantly stronger in bilinguals. The authors interpret these findings in the light of the concreteness effect: encountering concrete words results in richer activation of the semantic network compared to abstract words. They then combine this with a proposed view about the nature of bilingual semantic network, according to which there is more overlap between L1 and L2 concepts that are concrete (e.g. De Groot & Poot, 1997) than between abstract concepts. This would result in richer activation of the bilingual semantic network especially when encountering concrete words, which in turn would provide more stable grounds for learning. Thus, the authors suggest that bilingual benefit in novel word acquisition is at least partly based on greater sensitivity to semantic information. A similar conclusion was drawn in a study by Elgort et al. 2015 described earlier. They found larger semantic priming effect to novel L2 words in more proficient L2 learners compared to less proficient L2 learners. Elgort et al. (2015) suggest that the basis for proficient L2 learners' advantage in semantic learning might be a richer semantic network with ample lexico-semantic connections that allow efficient incremental learning. It is worth noting that Elgort and colleagues address a question of language proficiency in a within-language setting (richer semantic network in L2 facilitates L2 word learning), whereas Kaushanskaya and Rechtzigel suggest general word learning benefits (the nature of the bilingual semantic network supports more efficient word learning in general).

To conclude, in search for the underlying mechanisms for bilingual benefit in word learning, the notion of the facilitating effects of a rich semantic network on word learning has been suggested, and it is comparable to the previously proposed idea that a richer phonological network is a basis for more efficient novel word form acquisition in bilinguals. The two are not mutually exclusive and there is some evidence to support both conjectures. Regardless of the underlying mechanisms, the effects of language proficiency in novel word acquisition are far from being thoroughly explored. For instance, apart from Elgort et al. (2015), the studies reviewed above do not address the effects of language proficiency in lexicalization of novel

words or contextual word learning. These gaps in research in part motivated the current study, which will be described next.

1.4 The current study

The aim of the current study is to explore the nature of lexical representations of novel words learnt from context in more detail. The literature reviewed above shows that learning novel words in visually presented sentence contexts results in lexicalization of these words. The first goal of the current study is to look at contextual learning in the auditory domain, a neglected area in contextual learning research. On one hand, this involves testing the explicit knowledge of the nearly learnt words: how accurately the novel word forms can be recognized and how accurately the newly learnt word forms can be paired with their corresponding meaning. On the other hand, this involves testing the implicit knowledge of the newly learnt words: whether there is evidence of phonological and semantic lexicalization of the novel words. To this end, the experiment conducted consists of two testing sessions: the initial learning phase with immediate testing of the novel word acquisition, followed by a second testing session 48 hours later. The second goal of the current study is to investigate the effects of language proficiency on the different aspects of novel word acquisition: whether higher L1 or L2 proficiency facilitates the process of novel word acquisition in the aforementioned dimensions (i.e. explicit and implicit knowledge of word forms and word meanings). Vast majority of the reviewed studies investigating language proficiency effects on vocabulary acquisition used explicit measures of learning. Thus, the current study takes a step further by investigating language proficiency effects during lexicalization of novel words. Additionally, although previous studies investigating the effects of language proficiency generally show a benefit of higher proficiency in vocabulary acquisition, which aspect of vocabulary acquisition is facilitated by higher language proficiency is still unclear. Does this benefit only manifest in within language learning (i.e. L1 proficiency facilitates L1 vocabulary learning) or does high language proficiency aid word learning across langauges as well (i.e. L1 proficiency facilitates L2 vocabulary learning or L2 proficiency facilitates L1 vocabulary learning)? Previous studies mostly speak to the former (i.e. within language benefit), whereas in the current study, the word learning task is a simulation of

learning synonyms in one's native language, thus aiming to answer the question of whether L2 proficiency facilitates L1 word learning (and whether L1 proficiency facilitates L1 word learning). Learning L1-like words was also chosen in order to minimize the effects of potential L2 proficiency advantage in learning phonotactically unfamiliar word forms. As both word form and meaning acquisition are of interest, but the two are difficult to tease apart (probing meaning acquisition always has an element of recognizing the word form linked to it), native language word forms were considered the most neutral option for revealing potential language proficiency effects in meaning acquisition.

The meaning acquisition is investigated utilizing a semantic priming paradigm, where a novel word is paired with its meaning (related condition) or with a meaning of another novel word (unrelated condition). In addition to prime-target pairs of the novel word and its meaning (e.g. *cathedruke – basket*), a condition where the novel word is paired with a semantic associate of its meaning is included (e.g. *cathedruke – weave*). This semantic associate condition provides a crucial addition to previous designs, because - as pointed out by Tamminen & Gaskell (2013) - priming effects between a word form and its meaning might only demonstrate an isolated association between the two, i.e. the influence of the novel item might not extend to a broader semantic network. This is true especially when the learning phase involves explicit presentation of the meaning assigned to the novel word. If, however, a priming effect is found between a novel word form and its semantic associate (utilizing semantic associates not presented during the learning phase), this suggests that the novel word has the ability to activate other semantically related items in the mental lexicon, i.e. the novel item has been integrated into the semantic network. A demonstration of semantic learning going above and beyond associative learning in written novel word acquisition was reported by Tamminen & Gaskell (2013). In their experiments, learning novel words (e.g. *feckton*) paired with meanings (e.g. *feckton is a type of cat that has stripes and is bluish-grey*) was tested by lexical decision task, utilizing masked and unmasked priming. The targets in this task were semantic associates of the learnt meanings, e.g. *dog, mouse* and *kitten* for a meaning *cat*. For unmasked primes, the priming effect was found both immediately and a week after learning the novel words, but only a week later in the masked priming test. Additionally, Clay, Bowers, Davis & Hanley (2007)

provide evidence of semantic learning beyond associative learning of written words. In their study, a form of Stroop effect, a picture-naming interference, served as a measure of automatic, semantic processing of newly learnt words. The novel word forms were learnt paired with a definition and a picture and the learning test administered involved picture naming while ignoring a distractor word presented simultaneously. The authors point out that this task is less sensitive to associative relations than priming tasks are and thus serves to probe semantic rather than associative learning. What was found was that pictures presented with newly learnt words that matched their semantic category were named slower than pictures presented with unrelated novel words. This effect was not seen immediately, but a week after learning the novel words. Finally, the study by van der Ven and colleagues (2015) described earlier reported semantic priming effect between novel words and their semantic associates only 24 hours after the learning phase. The current study adds to these findings of semantic learning of novel words by investigating spoken word rather than written word acquisition and by testing the emergence of the lexicalization effect 48 hours rather than a week (or 24 hours) after the learning of the novel words. Importantly, the contextual word learning task utilized in the current study further ensures that the learnt novel items are not represented as mere isolated, episodic associations, as the novel word form is never explicitly paired with its meaning during the learning task. Furthermore, Frishkoff et al., 2010 used near-synonyms of the novel words as targets in the priming task, whereas the target words in the current study are never near-synonyms of the novel word meanings, thus requiring more extensive activation of the semantic network before the priming effects can be detected. As such, the more stringent test of semantic learning employed in the current study has not been used in contextual word learning studies in auditory modality before.

Finally, the current study utilizes more objective measures of language proficiency compared to several previous studies in the area of bilingualism, where self-assessments have been used (e.g. Kaushanskaya & Marian 2009, Nair et al. 2015). The correlation of self-assessment measures and performance based language tests varies. Learners with higher proficiency levels tend to underestimate their proficiency whereas learners with low proficiency typically overestimate their abilities (e.g. Alavi & Akbarian 2008; Edele, Seuring, Kristen & Stanat, 2015).

Therefore, the level of language proficiency in the current study is operationalized as a vocabulary test score (Izura, Cuetos, & Brysbaert, 2014; Paul Nation's Vocabulary Size Test, retrieved from https://www.victoria.ac.nz/lals/about/staff/paul-nation#vocab-tests).

In sum, the current study aims to answer the following research questions:

1. Can novel word acquisition take place via contextual inference from aurally presented sentences?
2. Does contextual novel word acquisition in auditory domain result in lexicalization of the novel words and if so, what is the timeline for this process?
3. Does acquisition of novel words from context vary as a function of first language or second language proficiency?
4. Does lexicalization of novel words learnt from context vary as a function of first or second language proficiency?

2. Methods

2.1 Participants

30 University students and volunteers (9 males) participated in the study. They were paid or received course credit for their participation. Each participant was a native English speaker with Spanish as their strongest second language. All participants had normal or corrected-to-normal hearing and vision. One participant was excluded from the analyses due to the experimenter's error in giving instructions, which resulted in considerably different task performance compared to the other participants. The data from the remaining 29 participants was used in the analyses. The average age of the participants was 23 years, ranging from 18 to 40. The average number of foreign languages known was 2.7, ranging from 1 to 5 and the average number of years spent studying L2 (Spanish) was 8.2, ranging from 0.6 (7 months) to 22 years. Most of the participants had learnt their L2 mainly via formal instruction, however, two participants had been exposed to their L2 for 18 years. Self-reported percentage of daily use of L2 was on average 11.5%, ranging from 0 to 50%. 79% of the participants estimated their use of L2 in daily life to be 20% or less (See table 1 for summary of participant characteristics).

Table 1. Participant characteristics (n = 29). Standard deviations are presented in parentheses after means.

	Age in Years	Education in Years	L2 Studying in Years	L1 vocabulary score (%)	L2 vocabulary score (%)	No. of Foreign Languages	Digit Span score	Matrices score (%)
Mean	23.3 (5.9)	16.6 (3)	8.2 (4.9)	86.4 (7.3)	32.5 (20.5)	2.7 (1.1)	6.7 (1.2)	84.8 (13.3)
Min	18	13	0.6	70	5	1	5	25
Max	40	25	22	99.5	75	5	9	100

2.2 The Experiment Design

Participants attended two sessions that were 48 hours apart. Each session took approximately 2 hours to complete. In the beginning of each session, an EEG cap was fitted and participants completed most of the tasks wearing it in a shielded room where the EEG recording for two of the tasks took place. Figure 1 shows the order of the tasks in each session.

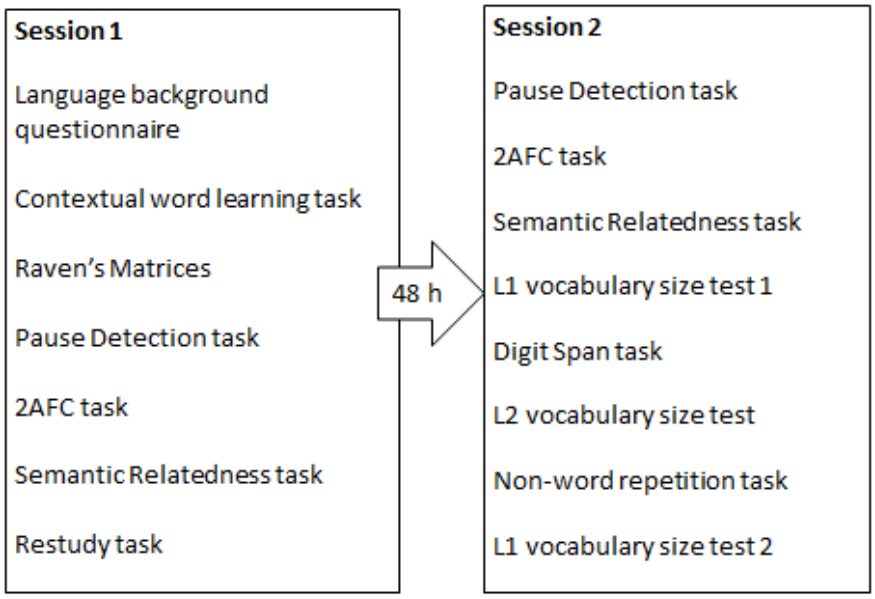


*Figure 1.*The order of the tasks completed in Session 1 and Session 2. L1 (English) vocabulary size test 1 (subset of Paul Nation's vocabulary size test version A) and L1 vocabulary size test 2 (online vocabulary test).

In session 1, participants first completed a language background questionnaire after which they completed a contextual word learning task that took approximately 30 minutes. In this task participants learnt novel words embedded in sentence contexts (learning phase). After the learning phase, participants completed a 5 minute distraction task (subset of Raven's Progressive Matrices), followed by three word learning tests designed to measure the initial learning of the novel words (Pause Detection, two-alternatives forced-choice (2AFC) and Semantic Relatedness judgments task). Finally participants completed a restudy task, where each of the novel words was presented one more time in a sentence context. In session 2, participants completed the same three word learning tests as in the first session, followed by 5 participant characteristics tasks that were designed to provide measures of first language (L1, English) and second language (L2, Spanish) vocabulary size and short-term memory (STM) capacity. These tasks were, in order of administration, English vocabulary size test 1 (a subset of Paul Nation's test version A), Digit Span test, Spanish vocabulary size test (Lextale-Esp by Izura, Cuetos & Brysbaert, 2014), Non-word Repetition test (Gathercole & Baddeley, 1994) and English vocabulary test 2 (online test). Each task is described in more detail below.

2.3 Tasks

2.3.1 General features of experimental tasks

All the experimental tasks were completed on a computer. The stimuli were presented and the responses recorded using Presentation ® software (Version 17.1, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com). Responses were always indicated by pressing one of two buttons (number 1 or number 2 on the right hand side of the keyboard) with options for each button shown on the screen.

2.3.2 Contextual novel word learning task

This task was designed to expose participants to novel words in auditory sentence contexts where the meaning of each novel word could be inferred from the context. The task thus served as the main learning phase before the learning of the novel words was tested. The participants were informed that their knowledge of the novel words will be tested later. There

was a short practice session before the actual task begun. Participants started each trial by pressing a button. After this, a fixation cross was presented on the screen during the time the sentences were played over the headphones. There were 32 novel words to learn. In each trial, participants heard three sentences over the headphones, each of which contained the same novel word (e.g. *When Aaron does laundry he carries his clothes downstairs in a cathedruke first* [cathedruke = basket in this context]). Each novel word was an English non-word (see Word Forms in Materials). After the sentence triplet, participants were presented with a question on the computer screen regarding the meaning of the novel word. They had to choose the right semantic category for the novel word out of two options (e.g. *A CATHEDRUKE is 1) a profession 2) a container*) by pressing number 1 or 2 on the right hand side of the keyboard. Each wrong option for any given word was a correct option for another novel word in the experiment. The correct option for each novel word appeared first in 50% of the questions. After answering the question, participants pressed Space bar to start the next trial.

Each novel word appeared in six different sentence contexts, divided in two different sentence triplets. The sentences in each triplet were in a fixed order so that the cloze probability of the sentences increased from the first to the third sentence in a triplet. The two triplets for any given novel word were presented in a randomized order with a restriction that there was a maximum of 14 triplets between the triplets of the same word. This restriction was implemented to ensure that the repetition of each novel word would happen relatively shortly after the first presentation. However, the two sentence triplets for the same novel word could have been presented directly following each other, but this was only possible for a maximum of 4 novel words out of 32. The sentence triplets were divided into 4 different blocks (each containing both triplets of 8 novel words), the block order of which was rotated across participants. The interstimulus interval (ISI) of the sentences in a triplet varied between 1000, 1500 and 2000 ms.

The completion of the task took approximately 30 minutes. At the end of the task, participants had been exposed to each novel word 6 times aurally in the sentence contexts and twice

visually in the semantic category questions. EEG data was recorded simultaneously as participants completed the task, however, this data was not analyzed.[3]

2.3.3 Restudy of the novel words task

The purpose of the restudy task was to support the learning of the novel words by presenting them one more time in context and by narrowing down the possible meanings of the novel words with more restrictive questions about them. At the end of the first session, participants heard each novel word again in one of the sentence contexts they had already heard in the beginning of the session. After each sentence, the participants again chose the right semantic category for the novel word by pressing the number 1 or 2 on the right hand side of the keyboard. This time the questions were more restrictive (e.g. *A CATHEDRUKE is 1) a container that's waterproof, 2) a container that doesn't hold water*).

2.3.4 Pause Detection task

This task was designed to test whether the newly learnt word forms were lexicalized. Largely following the procedure used by Gaskell & Dumay (2003, Experiment 3), the participants were presented with words over the headphones and indicated with a button press whether there was a pause in the word they had just heard. Participants were instructed to respond as quickly as possible. Each trial started when the Space bar was pressed. There was a fixation cross during the time the word was presented, after which the options of 1) *Pause present* and 2) *Pause absent* were presented on the screen. Participants responded by pressing one of two buttons (number 1 or 2 on the right hand side of the key board).

Each participant was presented with 32 critical words (i.e. base words of the novel word forms that the participant had previously learnt in the contextual word learning task, e.g. *cathedral* for participants who had heard *cathedruke* in the contextual learning task), 32 control words (the base words for the other 32 novel word forms that the participant had not heard before,

[3] After fitting the EEG cap in the beginning of the testing sessions, the quality of the EEG recordings improved as the electrodes and the conductive gel settled. The EEG data was recorded during the contextual learning task (immediately after cap fitting), even though the data collected during this task was noisy and the focus of the EEG data analysis was in the data collected during the Semantic Relatedness task.

e.g. *boulevard* for participants who had not heard the novel word *boulevette* in the contextual learning task) and 64 fillers. Half of the critical, control and filler words presented in one testing session contained a pause. Each participant would hear the pause present version of any given base word in Session 1 and the pause absent version of this word in Session 2 or vice versa. The order of presenting the pause present and pause absent versions of any given base word was counterbalanced across participants, e.g. half of the participants assigned to list A would hear *cathe_dral* in Session1 and *cathedral* in Session2 (where _ indicates a pause) and the other half of the list A participants would hear these versions of the base word in the opposite order (*cathedral* in Session1, *cathe_dral* in Session2). The stimuli were presented in a fully randomized order with a unique randomization for each participant in each session. The duration of the stimuli (each aurally presented word) including fillers ranged from 410 to 1123 ms and was 746 ms on average. The critical and control words ranged from 487 to 1123 ms with a mean of 769 ms. Each critical-control word pair (e.g. *cathedral* in list A and *boulevard* in list B) were matched in length and uniqueness point as closely as possible (see Table 3 in Word forms), so that any differences in reaction times between critical and control words could be attributed to differences in resolving lexical competition due to an additional lexical competitor added for critical items but not for control items.

2.3.5 Two-Alternatives Forced-Choice (2AFC) task

The 2AFC task was designed to test the explicit recognition accuracy of the novel word forms. In this task, participants heard word pairs and their task was to decide which one of the two words was more familiar to them. Each trial started after the participant pressed the Space bar. After each word pair the options of 1) *firs word* and 2) *second word* were presented on the screen. The choices were indicated by pressing one of two buttons (number 1 or 2 on the right hand side of the key board). There was no time pressure for responses. Each of the 32 novel words (i.e. *cathedruke*) that the participants had learnt before in the contextual word learning task was played once followed or preceded by a foil of the given novel word (e.g. *cathedruke – cathedruce* or *cathedruce – cathedruke*). The order of the novel word - foil word pairs was fully randomized, as was the order of the words in each word pair (e.g. whether the novel word or the foil was played first). The second word was always played 1100 ms after the onset of the

first one, so that ISI of the word pairs varied between 171 ms and 623 ms. A unique randomization of the stimuli was presented for each participant. The duration of the stimuli (each word) ranged from 477 to 929 ms. The same task was completed in the second session with a new randomization for each participant.

2.3.6 Semantic Relatedness judgments task

The purpose of this task was to test the explicit knowledge and lexicalization of the novel word meanings learnt in the contextual word learning task. Response accuracies, reaction times (RTs) and EEG responses (ERPs) were collected during the task. Participants heard word pairs over the headphones and made speeded decisions as to whether the two words were semantically related or not. Each trial started after participants pressed the Space bar. There was a fixation cross on the screen during the time the word pair was played over the headphones. After this the options of 1) *Related* and 2) *Unrelated* were shown on the screen until the participants pressed on of the two buttons to indicate whether the words were related or unrelated (number 1 or 2 on the right hand side of the key board). The word pairs were either related or unrelated and they were presented in three different conditions (see Table 2 for an example).

Table 2. Example of related and unrelated trials in the Semantic Relatedness task. An example of the novel words Cathedruke and Molekyn in Real(R), Meaning (M) and Semantic Associate (SA) conditions for related and unrelated trials. The related targets of Cathedruke are used as unrelated targets for Molekyn and vice versa. Note that this is an example of how the related and unrelated targets between different items were swapped, but does not mean that each item's unrelated target always came from the same other item for each condition (i.e. two items were not paired and their related and unrelated targets swapped with each other as might appear from the example above).

Novel word: Cathedruke

	Related prime - target	Unrelated prime - target
R	Basket - WEAVE	Basket - TOOTH
M	Cathedruke - BASKET	Cathedruke - DENTIST
SA	Cathedruke - WEAVE	Cathedruke - TOOTH

Novel word: Molekyn

	Related prime - target	Unrelated prime - target
R	Dentist - TOOTH	Dentist - WEAVE
M	Molekyn - DENTIST	Molekyn - BASKET
SA	Molekyn - DENTIST	Molekyn - WEAVE

For related trials, the Meaning (M) condition was a novel word with its allocated meaning (e.g. *cathedruke* – *basket*). The Semantic Associate (SA) condition was the novel word paired with the semantic associate of its allocated meaning (e.g. *cathedruke* – *weave*). The Real (R) condition was the meaning (i.e. English equivalent) of the novel word and its semantic associate (e.g. *basket* – *weave*). The second word of a word pair (the target word) was always a real English word, so that the novel words presented in M and SA conditions were always the first word of the word pair (i.e. the prime). Each of the 32 novel words was presented in each condition both in a related trial (e.g. *cathedruke* – *basket* for the M condition) and an unrelated trial (e.g. *cathedruke* – *dentist* for the M condition). A target in an unrelated trial for any given word was always a target in a related trial for another word used in the experiment. As such, each novel word, each meaning and each semantic associate of a meaning (e.g. *cathedruke* – *basket* – *weave*) were presented 4 times throughout the task, ensuring equal number of exposure to each item. The stimuli were presented in 6 different blocks so that no item (either as a prime or as a target) appeared twice in the same block (e.g. *basket* did not appear as a target in M condition (*cathedruke* – *basket*) and as a prime in R condition (*basket* – *weave*) within the same block). The order of the blocks was rotated across participants and the presentation of word pairs in each block was fully randomized. Each participant completed the same task in the second session, with a new randomization and block order. EEG data was recorded in both testing sessions. The ISI of word pairs varied between 500 and 700 ms in 50 ms steps to avoid expectancy effects in the EEG responses (Min et al. 2008). The duration of the primes ranged from 399 to 929 ms. The duration of target words ranged between 399 and 722 ms and were 584 ms on average for M condition and between 381 and 810 ms with 567 ms average for R and SA conditions. The phonemic length of targets in related and unrelated trials was matched as closely as possible to ensure that any differences in reaction times to the targets could be attributed to the experimental manipulation of related or unrelated preceding primes (See Appendix A, Table A2).

2.3.7 Participant Characteristics tasks

The Participant Characteristics tasks were completed at the end of the second session in order to collect information about the L1 (English) vocabulary size, L2 (Spanish) vocabulary size and STM capacity of the participants. Each participant also completed a set D (12 problems) of Raven's Progressive Matrices (Raven, 1958) as a distraction task in the first session.

2.3.7.1 L1 (English) vocabulary tasks

Two frequency informed English vocabulary size tasks were used. Participants completed an online vocabulary size test (Retrieved from http://testyourvocab.com/) where they ticked all the words they knew at least one meaning for. The estimate for their vocabulary sized was calculated automatically, based on the lowest frequency words ticked as known. Note that there were no incorrect answers in this test, which means that guessing would lead to a higher score. Due to already long testing sessions, only half of Paul Nation's vocabulary size test (monolingual, version A. Retrieved from https://www.victoria.ac.nz/lals/about/staff/paul-nation#vocab-tests) was completed. Each participant answered the last 50 questions of this pen and paper task, where words with different frequency of occurrence are presented in an uninformative context. The best description of the word's meaning is chosen from 4 options. Although both tests were administered, the score from Paul Nation's vocabulary size test was used as this was a more objective measure.

2.3.7.2 L2 (Spanish) vocabulary task

Participants completed a Lextale-Esp (Izura, Cuetos & Brysbaert, 2014), a Spanish vocabulary task in which they saw a list of Spanish words intermixed with Spanish pseudowords. The task was to tick all the words the participants knew were Spanish. They were informed that it was important to tick only the words they were certain of, as ticking wrong words in the task results in score reduction (each correct response was worth 1 point and each incorrect response was worth -2 points).

2.3.7.3 Short-term Memory tasks

Each participant completed an aurally administered Digit Span task (e.g. Wechsler, 1997) and a non-word repetition task (Gathercole & Baddeley, 1996). The digit span task was computerized, administered using the PsychoPy software (Peirce, 2007). Each trial in this task consisted of participants hearing a sequence of digits over the headphones after which they typed in the digits in the order of presentation. Participants pressed the Space bar to continue to the next trial. The length of a sequence of digits increased from 3 up to 9 digits (in 1 digit steps) if participants made no errors. Each sequence length was repeated twice in two separate trials and if an error was made, another trial with a sequence of the same length was presented. If two consecutive trials with the same length of sequences were repeated incorrectly, the test ended. Digit span of a participant was the last sequence length they could repeat twice correctly. In the non-word repetition task, participants heard non-words over the headphones one at a time and they had to repeat each non-word immediately after hearing it. Verbal responses were recorded and a total of 40 non-words were presented. The accuracy of the verbal responses was scored by a native English speaker. The Digit span score was used as a measure in the data analyses, because it is argued to be more directly linked to novel word acquisition than non-word repetition (Gupta, 2003).

2.4 Materials

2.4.1 General properties of the stimuli

Two experimental lists (list A and list B) were created with 32 word form – word meaning pairs in each list. Each word form was an English non-word that was paired with a meaning, an existing English word (e.g. *cathedruke* – *basket*). These word form – word meaning pairs are referred to as novel words, as they were the stimuli to be learnt in the contextual word learning task described above.

All materials were recorded in a sound proof room, spoken by a male, native English speaker. The intensity of the resulting recordings was equalized to 70 dB and played over headphones on a comfortable level.

2.4.2 Word forms

For the purposes of testing the novel word form acquisition, triplets of base words, non-words and foils were created for the experiment. The base words were existing English words that the non-words deviated from at the final vowel (e.g. *cathedral* – *cathedruke*). These base words were used in the Pause Detection task described earlier. The non-words (referred to as novel word forms) were used in the Contextual word learning task and Semantic Relatedness task described earlier. An alternative version of each non-word was also created to be used as a foil in a 2AFC recognition task described earlier. These foils deviated from the non-words at the final consonant or consonant cluster (e.g. *cathedruke* – *cathedruce*, see Appendix A, Table A1 for the full list). The total of 64 base word – non-word – foil triplets were divided into two experimental lists (list A and list B) and each participant was assigned to one of these lists of 32 triplets.

2.4.2.1 Base word properties

The base words for each non-word (e.g. *cathedral* for *cathedruke*) across experimental lists were matched in frequency of occurrence (per million) and uniqueness point (UP) as closely as possible (See table 3 for summary and Appendix A, Table A1 for the full list of base words in list A and list B).

Table 3. Properties of the base words in lists A and B. Means of frequency of occurrence per million (Frequency), uniqueness point (Uniqueness) and phonemic length (Phonemes) of the base words are shown with standard deviations in parentheses.

List A	Frequency	Uniqueness	Phonemes
Min	2	3	6
Max	18	6	9
Mean	6.1 (4.4)	4.5 (0.8)	6.9 (0.9)

List B	Frequency	Uniqueness	Phonemes
Min	2	3	6
Max	19	6	8
Mean	6.0 (4.5)	4.5 (0.9)	6.7 (0.7)

The UP of each base word was defined by WebCelex (Retrieved from http://celex.mpi.nl/) database search as the first phoneme that made the word a unique sequence (i.e. no other word matched the beginning of the given word from this point on). For the total of 64 base

words, lemma frequency of occurrence per million (defined by WebCelex database search) ranged from 2 to 19 with a mean of 6.0. The length of the base words ranged from 6 to 9 with a mean of 6.8. The uniqueness point (UP) was between the 3rd and the 6th phoneme and was on average 4.5.

2.4.2.2 Non-word properties

The non-words were trisyllabic and bisyllabic words that followed the English phonotactic rules. Out of the total of 64 non-words, 31 were from the materials of Gaskell & Dumay (2003) and 33 were created following the same criteria as closely as possible. The length of the total of 64 non-words ranged from 6 to 10 phonemes with an average of 7.0. The UP ranged from the 5th to the 8th phoneme and was on average 5.9. Note that if the non-words were learnt, these UPs would also be the new UPs of the base words. There were 12 non-words (6 in each list) that had the same UP as their base words, which means that if learnt, these words would not move the UP towards the end of the base word. Although not optimal, these non-words would still become an additional phonological competitor in the process of word recognition (a new word that is indistinguishable from the base word all the way up until the UP of the base word).

2.4.2.3 Fillers

The filler words used in the Pause Detection task were English words with a frequency of occurrence (per million) ranging from 2 to 16 with an average of 5.7. The length in phonemes was in the range of 3-9 and 5.1 on average. There were 16 trisyllabic, 23 bisyllabic and 25 monosyllabic fillers. The same list of fillers was used for all the participants and in both testing sessions.

2.4.2.4 Pauses

For Pause Detection task, a 200 ms pause was inserted to the base words just before or at the UP. Where inserting the pause would distort the base word, a pause was inserted as late before the UP as possible. Pauses in filler words were inserted in the following way: for trisyllabic

words after the first syllable, for bisyllabic words just before the first vowel or just after the last vowel and for monosyllabic words just before or after the vowel.

2.4.3 Word Meanings

Each non-word from both experimental lists (A and B) had a meaning (an English word) assigned to it. Hence, there were 32 meanings in total, one non-word from each list assigned to one meaning (e.g. *cathedruke* from list A and *boulevett* from list B both assigned to a meaning *basket*). The meanings of the non-words were highly concrete, imaginable English words with imaginability ratings based on web interface of MRC Psycholinguistic database (Wilson, 1988) ranging from 483 to 645 (out of maximum of 700) with an average of 593.8. Four of the meanings (*bicycle, casino, airport* and *finger*) did not have an imaginability rating available in the database. The frequency of occurrence of the meanings was on average 55.9 and ranged from 4 to 143.

2.4.4 Sentence contexts

The sentence contexts and novel words (novel word here is a non-word – meaning pair) used in the final experiment were selected from a larger pre-tested sample of sentence contexts and novel words. The larger set of eight sentences for 48 novel words were created and tested with a sentence completion task. This task was completed by 26 participants who did not take part in the final experiment. The participants were native English speakers and filled in the sentence completion task online, using Qualtrics online survey program (Qualtrics, Provo, UT, USA). In this task the participants filled in the first word they thought of when provided with the sentences that were missing the final word (e.g. *When Aaron does laundry he carries his clothes downstairs in a* _____.). The sentences with the most consistent completion, i.e. the highest cloze probability (the percentage of responses with the same word, *basket* in the example above) were chosen for the final set of 6 sentences for each of the final 32 word meanings. The mean cloze probability for the final set of sentences was 91.5% (See Appendix A, Table A3 for more details).

As explained in more detail in the description of the Semantic Relatedness task, each novel word had a meaning assigned to them, as well as a semantic associate to the meaning (e.g. meaning: *basket*; semantic associate: *weave*). Percentage of the sentences that could be intelligible using either the meaning or the semantic associate of the novel word was 21 % of all sentences. Only one novel word (*Airport*) had a 100% replaceability, as all 6 sentences containing the novel word could have been intelligible using either the meaning (*airport*) or its semantic associate (*terminal*). Even though the questions in the Restudy task were more restrictive, 22% of the questions could still be answered correctly using the semantic associate instead of the meaning of the novel word (See Appendix A, Table A3). As the meanings the participants gave for each novel word were not explicitly tested, it is important to ensure that the right meaning was inferred in the learning phase. The most confusable novel words are expected to be the ones with 50% or more replaceability in the learning phase (i.e. at least 3 out of 6 sentence contexts would make sense using the semantic associate of the novel word meaning instead of the meaning itself) combined with replaceability in the restudy phase. There were only four novel words that fit this criterion (*airport, hotel, map* and *pocket*). It is hence reasonable to assume that the combination of the learning phase and the restudy phase did not encourage learning of the semantic associates of the novel word meanings, but rather the actual meanings of the novel words.

3. Results

3.1 Analyses of Response Accuracies

The focus of these analyses was on whether the recognition accuracies in the word learning tests were above chance level and whether the level of recognition accuracies changed between the testing sessions (Session 1 was the same session where the novel words were learnt and Session 2 was 48 hours after the Session 1). The relationship between performance in the word learning tests and language proficiency and short-term memory (STM) capacity were also investigated.

3.1.1 2AFC task

3.1.1.1 Response accuracy

The mean percentage of correct responses (%Correct) was investigated utilizing one-sample t-test (2-tailed) for both testing sessions separately. The mean %Correct score was compared to a chance level mean performance of 50%. Here and in the following results, the data from all 29 participants was used unless otherwise specified.

For session 1, the mean %Correct score (M = 85.99, SD = 9.62) was significantly higher than chance level performance ($t(28) = 20.15$, $p < .001$). For session 2, the mean %Correct score (M = 88.8, SD = 7.59) was also reliably higher than chance level ($t(28) = 27.54$, $p < .001$). This suggests that participants could recognize the correct word forms immediately[4] after learning them and after a 2 day delay.

3.1.1.2 Change in response accuracy between sessions

The change in the percentage of correct responses (%Correct) between sessions was inspected. Subtracting %Correct1 (Session 1) score from %Correct2 (Session 2) score gave a value that

[4] In the results section, immediate performance refers to the test performance in the same session where the novel words were learnt (Session 1), but none of the word learning tests were administered immediately after the learning phase (contextual novel word learning task), but after at least one 5-minute distraction task (Raven's Matrices). The same session test performance is described as immediate for brevity.

indicates the direction of change in performance for each participant (Difference score). One-sample t-test (2-tailed) was run comparing mean Difference score to a mean of 0, corresponding to no change in %Correct score between the two sessions. The Difference score was positive for 58.6% of the participants (17 out of 29), negative for 24.1% of the participants (7 out of 29) and zero for the remaining 17.2% of the participants (5 out of 29). The mean Difference score ($M = 2.8$, $SD = 7.2$) differed reliably from 0 ($t(28) = 2.1$, $p < .05$), suggesting that on a group level the participants' recognition of the novel word forms improved reliably from Session 1 to Session 2.

3.1.1.3 Predictors for response accuracy

In order to investigate participant characteristics associated to the word form recognition accuracy, a multiple linear regression analysis was run for both Session 1 and Session 2 separately. For each analysis, the percentage of correct responses (%Correct) in the given session was the dependent variable, whereas the predictors in Session 1 were First language score (L1), Second language score (L2) and Short-term memory capacity score measured with Digit Span task (STM). For Session 2, the same predictors were used in addition to %Correct in Session 1 (%Correct1). See Table 4 for correlations between the candidate variables. Model building was carried out utilizing a backward elimination procedure. Multicollinearity diagnostics showed that tolerance and variance inflation factor (VIF) for each predictor was within acceptable range (tolerance ≥ .2 and VIF ≤ 5), following commonly accepted cutoffs. Multicollinearity was also inspected using model dependent VIF cutoff values (see Craney & Surles, 2002). This procedure allows detection of predictors for which other predictors in the model have more explanatory power than this predictor has for the dependent variable. No such predictors were found, and thus multicollinearity is not considered an issue in the models.

Table 4. Correlations of candidate variables in 2AFC task regression models. Pearson, 2-tailed. %Correct1 and %Correct2 (response accuracies in Session1 and Session2, respectively), L1 and L2 (L1 and L2 vocabulary score), STM (digit span score).

	%Correct1	%Correct2	L1	L2
%Correct2	0.67 ***			
L1	0.42 *	0.5 **		
L2	0.09	0.42 *	0.34 .	
STM	0.27	0.32	0.08	0.20

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

3.1.1.3.1 Session1

The only significant predictor in the final model was L1 (R^2 =.17, F(1,27) = 5.65, p < .05), explaining 17% of the variance in %Correct in Session 1. This suggests that higher level of L1 proficiency is associated with higher immediate recognition accuracy of novel word forms.

3.1.1.3.2 Session2

The final model with %Correct1 and L2 as predictors explains 55% of the variance in %Correct2 (R^2 = .58, Adjusted R^2 =.55, F(2,26) = 18.18, p < .001). %Correct1 was the strongest predictor (B = .5, ß =.64, t(26) = 5.02, p < .001) followed by L2 (B = .13, ß = .36, t(26) = 2.85, p < .01). Uniquely %Correct1 explains 40% and L2 explains 13% of the variance in %Correct2 score.[5] These results suggest that higher recognition accuracy immediately after learning the novel word forms along with higher L2 proficiency are associated with higher recognition accuracy of the novel word forms after a 2 day delay. The results can also be interpreted as higher L2 proficiency being linked to higher delayed recognition accuracy after controlling for immediate recognition accuracy. As such, the results reflect an association between L2 proficiency in %Correct gains over time. In order to explore the absolute level of recognition accuracy in Session 2, instead of the gains in recognition accuracy over time, immediate recognition accuracy (%Correct1) was

[5] Unique percentage of variance explained by a predictor is calculated by squaring the semipartial correlations of each predictor (Tabachnick & Fidell, 2001).

excluded from the set of candidate variables and the model selection was carried out again. This way the final model had L1 as a single predictor, explaining 25% of the variance in delayed recognition accuracy (R^2 = .25, F(1,27) = 9.15, p < .01).

3.1.1.4 Summary

Overall, participants identified the correct novel word forms successfully, as indicated by above chance level response accuracy in both sessions. The recognition of the novel word forms also improved after a 2 day delay. The multiple linear regression analyses showed that immediate recognition performance of novel word forms strongly predicts the delayed recognition success of these words. In addition, L1 proficiency was important in recognition of novel word forms immediately after learning them as well as after a delay, whereas L2 proficiency was linked to increase in recognition accuracy over time. These results will be considered further in the Discussion.

3.1.2 Semantic Relatedness task

3.1.2.1 Response accuracy

The mean percentage of correct responses (%Correct) was investigated utilizing one-sample t-test (2-tailed) for both sessions and each condition (Real word (R), Meaning (M) and Semantic Associate (SA)) separately. The mean %Correct score was compared to a chance level performance (mean of 50%). For session 1, the mean %Correct scores (R condition: M = 96.12, SD = 2.73; M condition: M = 77.37, SD = 12.82; SA condition: M = 71.61, SD = 12.39) were significantly different from chance level (R condition: t(28) = 91.1, p < .001; M condition: t(28) = 11.5, p < .001; SA condition: t(28) = 9.4, p < .001) in all three conditions. For session 2, the mean %Correct scores (R condition: M = 96.77, SD = 2.13; M condition: M = 80.33, SD = 12.89; SA condition: M = 76.35, SD = 13.75) also differed reliably from chance (R condition: t(28) = 118.43, p < .001; M condition: t(28) = 12.68, p < .001; SA condition: t(28) = 10.32, p < .001) in all three conditions. P-values were Bonferroni-corrected for 6 tests. These results suggest that participants could successfully make judgments of the word meanings between two known words (R condition) as well as between word pairs that consisted of a novel and a known word

(M and SA condition). This semantic judgment performance was reliably above chance level both immediately after learning the novel word meanings and after a 2 day delay.

3.1.2.2 Change in response accuracy between sessions

As above in the results for the 2AFC task, difference scores were used to inspect the change in %Correct between sessions in each condition (R, M and SA). The percentage of participants with a negative , positive and 0 Difference scores are shown in Table 5.

Table 5. Difference scores in Semantic Relatedness task. Percentage of participants with negative, zero or positive Difference score (Response accuracy in Session 1 subtracted from response accuracy in Session 2) for Real (R), Meaning (M) and Semantic Associate (SA) condition.

	Percentage of participants with Diff. score		
Condition	< 0	= 0	> 0
R	27.6	34.5	37.9
M	17.7	17.2	65.5
SA	13.8	13.8	72.4

In the R condition, the mean Difference score (M = .45, SD = 2.70) did not differ reliably from 0 ($t(28) = 1.30$, $p = .21$). In M condition, however, the mean Difference score (M = 2.96, SD = 5.72) was significantly higher than 0 ($t(28) = 2.79$, $p < .01$). This was also true for the mean Difference score in SA condition (M = 4.74, SD = 5.8) compared to 0 ($t(28) = 4.40$, $p < .001$). This pattern of results suggests that the accuracy in which participants made semantic judgments between two known words (R condition) didn't improve over time, whereas semantic judgments between novel and known words (M and SA conditions) were more accurate after a delay than immediately after learning the novel words (see Figure 2).

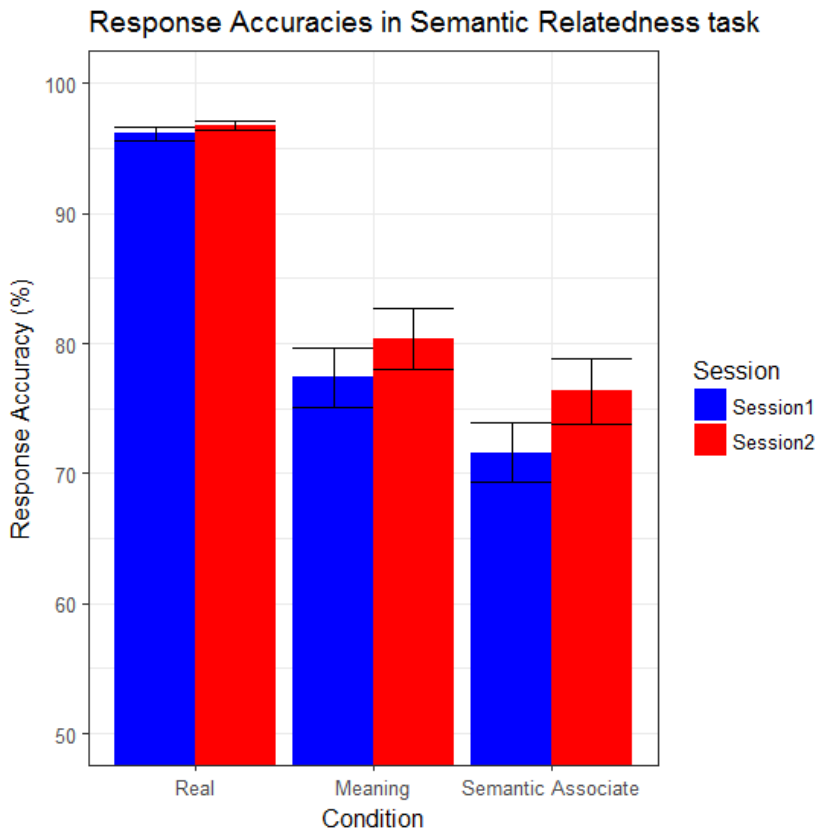


*Figure 2.*Response accuracies in the Semantic Relatedness task. Response accuracies by Condition and Session, starting at 50%. Error bars are standard errors.

3.1.2.3 Predictors for response accuracy

Multiple regression analysis was run for each of the three conditions (Real word (R), Meaning (M) and Semantic Associate (SA)) for both sessions separately with the percentage of correct responses (%Correct) as a dependent variable. The set of candidate predictors for each of these analyses were as described above in the results for 2AFC task. The same model dependent VIF cutoff screening was used and it was confirmed that multicollinearity was not an issue in the final models. (See Table 6 for correlations between candidate variables for the models). Because %Correct1 as a predictor for %Correct2 reflects the change in response accuracy between the testing sessions (i.e. Session 2 performance relative to Session 1 performance), the analyses of performance in Session 2 were also carried out without %Correct1 to discover the best predictors for the absolute performance level in Session 2.

Table 6. Correlations of candidate variables in Semantic Relatedness task regression models. Pearson, 2-tailed. R1 and R2 (response accuracies for Real word condition in Session1 and Session2, respectively), M1 and M2 (response accuracies for Meaning condition in Session1 and Session2), SA1 and SA2 (response accuracies for Semantic Associate condition in Session1 and Session2). L1 and L2 (L1 and L2 vocabulary score), STM (Digit span score).

	R1	M1	SA1	R2	M2	SA2	L1	L2
M1	0.25							
SA1	0.28	0.84 ***						
R2	0.41 *	0.41 *	0.42 *					
M2	0.4 *	0.9 ***	0.86 ***	0.48 **				
SA2	0.39 *	0.85 ***	0.91 ***	0.47 *	0.93 ***			
L1	0.35 .	0.45 *	0.56 **	0.36 .	0.55 **	0.51 **		
L2	0.11	0.30	0.47 *	-0.01	0.35 .	0.38 *	0.34 .	
STM	0.15	0.48 **	0.59 **	0.22	0.54 **	0.6 **	0.08	0.20

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

P-values were Bonferroni-corrected for 20 tests (i.e. for each global F-test of the final models and each t-test of each predictor in the final models) and these corrected p-values are reported unless otherwise mentioned.

3.1.2.3.1 Session 1

Real word Condition: None of the available predictors reliably explain the variance in %Correct score in the Real word condition[6].

Meaning Condition: Before correction for multiple comparisons, the final model with STM and L1 as predictors explains 35% of the variance in %Correct score (R^2 = .40, Adjusted R^2 =.35, F(2,26) = 8.64, p < .01). STM score was the strongest predictor (B = 4.62, ß = .44, t(26) = 2.92, p < .01) followed by L1 (B = .72, ß = .42, t(26) = 2.73, p < .05). However, after p-value correction for multiple comparisons, neither of the predictors reach significance at alpha level .05,

[6] None of the predictors reach statistical significance at alpha level .05 even *before* p-value correction for multiple comparisons.

although the model as a whole does. These results suggest that none of the available predictors reliably explain variance in %Correct score, although higher level of L1 and STM capacity might be associated with higher accuracy in semantic judgments between novel words and their meanings.

Semantic Associate Condition: The final two predictor model of STM and L1 (R^2 = .61, Adjusted R^2 =.58, F(2,26) = 19.31, p < .001) explains 58% of the variance in %Correct score. STM was the strongest predictor (B = 4.62, ß = .55, t(26) = 4.36, p < .01), followed by L1 (B = .72, ß = .51, t(26) = 4.08, p < .01). Uniquely STM explains 30% and L1 26% of the variance in %Correct score. This suggests that a higher level of STM capacity and L1 are associated with higher accuracy in semantic judgments between novel words and their semantic associates.

3.1.2.3.2 Session 2

Real word condition: Before correction for multiple comparisons, the final model with %Correct1 as the only significant predictor explains 16% of the variance in %Correct score in Session 2 (R^2=.17, F(1,27) = 5.35, p < .05). However, after p-value correction this model is no longer reliable. In order to investigate the predictors for the absolute performance level in Session 2, an analysis was carried out with %Correct1 excluded from the variable list. No reliable predictors were found. The results suggest that none of the available predictors explain the performance in Session 2, although higher accuracy in Session 1 is potentially associated with higher performance in Session 2.

Meaning condition: Before correction for multiple comparisons, the final model with %Correct1, STM and L1 as the predictors (R^2=.86, Adjusted R^2= .85 , F(3,25) = 52.17, p < .001) explains 85% of the variance in %Correct score in Session 2. The strongest predictor was %Correct1 (B = .72, ß = .72, t(25) = 7.5, p < .001), followed by STM (B = 1.9, ß = .18, t(25) = 2.1, p < .05) and L1 (B = .37, ß = .21, t(25) = 2.49, p < .05). However, after correction for multiple comparisons, only %Correct remains significant at alpha level .05 (p <.001). This suggests that higher accuracy in semantic judgments between a novel word and its meaning immediately after learning them is associated with higher accuracy of semantic judgments for these words after a 2 day delay. Analysis run with %Correct1 excluded from the variable list led to a final

model of L1 and STM as predictors (R^2 = .55, Adjusted R^2 =.52, F(2,26) = 16, p < .001), explaining 52% of the variance in %Correct2. L1 was the strongest predictor (B = .89, ß = .51, t(26) = 3.87, p < .05) followed by STM (B = 5.24, ß = .50, t(26) = 3.8, p < .05). Uniquely L1 explains 26% and STM explains 25% of the variance in response accuracy in Session 2.

Semantic Associate condition: The final model with %Correct1 as the only significant predictor (R^2= .82, F(1,27) = 124.3, p < .001) explains 82% of the variance in %Correct score in Session 2, suggesting that higher accuracy in semantic judgments between a novel word and its semantic associate immediately after learning the meaning of the novel word is associated with higher accuracy of these semantic judgments after a 2 day delay. Analysis run without %Correct1 as a predictor led to a final model of STM and L1 as predictors (R^2 = .57, Adjusted R^2 =.54, F(2,26) = 17.16, p < .001), explaining 54% of the variance in %Correct2. STM was the strongest predictor (B = 6.26, ß = .56, t(26) = 4.35, p < .01) followed by L1 (B = .86, ß = .46, t(26) = 3.57, p < .05). Uniquely STM explains 31% and L1 explains 21% of the variance in response accuracy in Session 2.

3.1.2.4 Summary

Accuracy of semantic judgments between word pairs in all three conditions were reliably above chance level both immediately after learning novel words and after a 2 day delay. In both novel word conditions (M and SA conditions), there was a statistically reliable increase in performance accuracy between the two sessions, whereas the performance accuracy for semantic judgments between two known words (R condition) did not increase reliably between sessions.

Multiple linear regression analyses showed that L1 proficiency and STM capacity play a role in immediate semantic judgment performance that requires linking the new word form to its semantic field (SA condition). These predictors might also be associated to immediate performance for novel word meanings (M condition), although these associations were not reliable. The strongest predictor for the performance in Session 2 in the novel word conditions was the performance in the first session. Additionally, L1 proficiency and STM capacity were

associated to the absolute delayed performance in the novel word conditions. None of the predictors have a significant effect on the performance in the R condition, although the immediate performance might be associated to the performance after a 2 day delay. These results will be considered further in the Discussion.

3.2 Analyses of Reaction Times and Event Related Potentials

3.2.1 Model Selection Procedure

Linear mixed effects (LME) regression analyses were applied to the Reaction time (RT) data in Pause Detection task, RT data in Semantic Relatedness task and the Event Related Potentials (ERP) data in the Semantic Relatedness task. The LME approach was chosen for the current investigations because it allows the use of continuous predictors, it can provide more accurate parameter estimates as individual variation of random effects (e.g. subjects and items used in the study) is taken into account and because it handles missing data appropriately (Baayen, Davidson & Bates, 2008).

A model selection procedure which allowed simultaneous comparison of several nested and non-nested candidate models was used. Model building was carried out incrementally by first adding individual predictors and subsequently, their interactions. Starting from a 0-model with no predictors, variables were added one at a time until a more complex model no longer improved the model fit compared to the preceding, simpler model. The primary interest variables Condition, Session and Trial type were added first and the secondary interest variables L1 and L2 proficiency after that. The addition of variables was done in parallel for different combinations of the candidate variables (e.g. adding Condition or Session or Trial type to a 0-model with no predictors and comparing the model fit of these three one-predictor models). The model choice procedure was based on Akaike Information Criteria (AIC), a measure of model fit enabling comparison of nested and non-nested models (Akaike, 1985). For small sample sizes and moderate number of parameters in a model, a correction of AIC, known as AICc, is recommended to avoid over-fitting of the models. In the current model selection this

correction was utilized and the final model was chosen based on the lowest AICc value. The crossed random effects Subject and Item were also chosen based on AICc values.

For the RT data in the Pause Detection task, the primary interest variables in model selection were Condition (Control, Critical), Session (Session1, Session2) and Trial type (Pause present, Pause absent). For the RT and ERP data in the Semantic relatedness task, the primary interest variables were Condition (Real word (R), Meaning (M), Semantic Associate (SA)), Session (Session1, Session2) and Trial type (Related, Unrelated).

3.2.1.1 Model fitting procedure

Analyses were run on R software (R Core Team, 2017), utilizing the lmerTest package (Kuznetsova, Brockhoff & Christensen, 2017). The Candidate models were fitted using maximum likelihood (ML) estimates, which is a recommended practice for model selection when the candidate models have a different fixed effects structure (Bates, 2010). The final models were fitted using restricted maximum likelihood (REML) and p-values were calculated based on Satterthwaite's approximation for degrees of freedom. Simulations comparing different methods for evaluating significance of fixed effects in mixed effect models have demonstrated that Satterthwaite's approximation for models fitted with REML produce acceptable Type I error rates, even for small sample sizes (Luke, 2016). Note that the coefficient estimates for all the LME models are reported as unstandardized coefficients (coef. B in tables).

3.2.2 Reaction Times in Pause Detection task

Out of all the trials from 29 participants (3712 observations), the total of 3.7% of the trials was lost due to incorrect responses or subject specific outliers (defined as 3 standard deviations from the mean of a participant's reaction times (RTs) in any given trialtype-condition combination). Out of all the trials, 2.1% was lost to incorrect responses and 1.6% to outliers. As reaction times showed marked non-normality, an inverse transformation was applied to the data. Thus the dependent variable in the analysis was -1/RT instead of raw values of RTs. Note that the direction of inverse transformed -1/RT values (1/RT multiplied by -1) can be

interpreted in the same way as raw values of RTs, e.g. larger values represent slower responses (Kliegl, Masson & Richter, 2010).

The RT data with a total of 3575 observations was analyzed with linear mixed effects (LME) regression analysis. Treatment coding (also known as dummy coding) was used for categorical variables with Control condition, Session 1 and Pause absent trials as reference levels.

3.2.2.1 Final model

The final model had Trial type as the only fixed effect and Subject and Item as random effects. The R formula for the model is as follows:

Model1: -1/RT ~ TrialType + (1|Sub) + (1|Item)

The simple effect of Trial type showed that the transformed RTs were significantly longer for items with a pause present than items without a pause (B = 2.11E-04, $p < .001$). The RT data thus didn't show the learning effect as indexed by longer RTs to critical trials compared to control trials. The observed RTs showed a slight numerical difference between the trial types in the pause present trials only so that the RTs to critical trials were longer than responses to control trials. This difference was 3 ms in Session1 and 11 ms in Session2 (see Figure 3). The pooled RTs across pause absent and pause present trials showed a pattern with responses to critical trials 17 ms shorter than responses to control trials in Session1 and responses to critical trials 4 ms longer than responses to control trials in Session 2. These results are considered further in the Discussion.

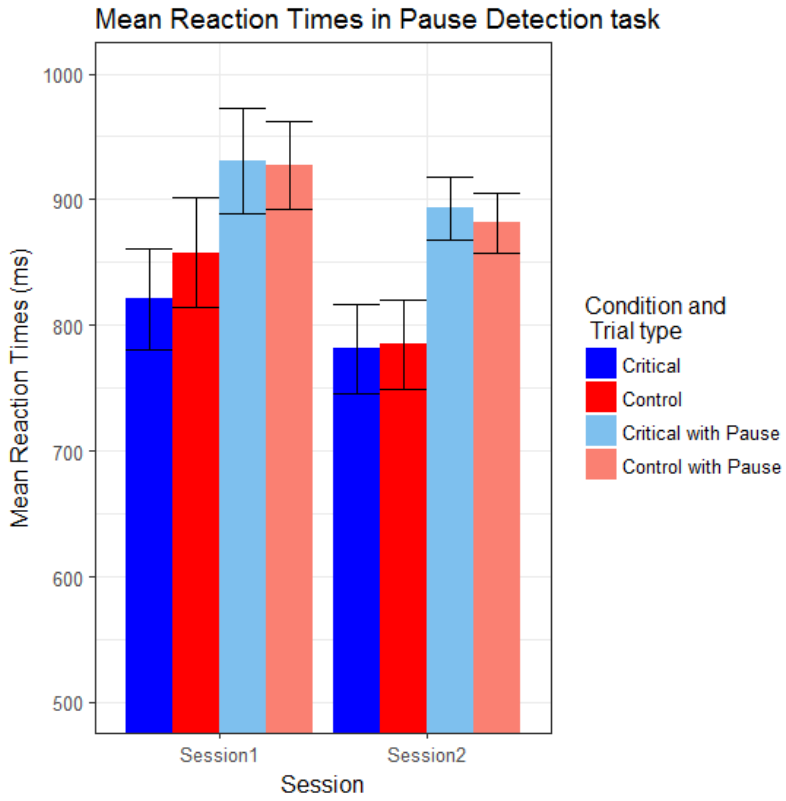


Figure 3. Reaction times in Pause Detection task. Observed values for reaction times by Condition and Trial type in Session 1 and Session 2, starting from 500 ms. Error bars are standard errors.

3.2.3 Reaction Times in Semantic Relatedness Task

Out of all the trials from 29 participants (11136 observations), the total of 22.58% of the trials were removed as incorrect responses (16.8% of total trials) or outliers (defined as 3 standard deviations from the mean of a participant's reaction times (RTs) in any given trialtype-condition combination (1.8% of total trials)) or responses longer than an absolute cut-off of 3000 ms (4.3% of total trials). As the number of correct trials for some participants in the novel word conditions (M and SA conditions) was low for related trials in particular(See Appendix B), the absolute cut-off was deemed necessary, as outliers removed based on measures of central tendency would not have been reliable. On the other hand, removing participants with too few trials left was not considered appropriate either, as this would potentially distort the results: participants with very few correct responses were probably the ones who found the task the most difficult. Removing them would have resulted in over-representation of RT data from

participants who performed well in the task. The final percentage of trials lost per condition in Session1 was 7.5%, 30.7% and 38.4% for real word (R), meaning (M) and Semantic Associate (SA) conditions, respectively. For Session2, the percentage of lost trials was 5.8%, 23.8% and 29.3% for R, M and SA conditions, respectively. As reaction times showed marked non-normality, an inverse transformation was applied to the data. Thus the dependent variable in the analysis was -1/RT instead of raw values of RTs. The RT data with a total of 8621 observations was analyzed with LME regression analysis. Treatment coding was used for categorical variables with R (Real word) condition, Session 1 and Related trials as reference levels.

3.2.3.1 Final model

The final model had Condition and Session as fixed effects and Subject and Item as random effects. The R formula for the model is as follows:

Model1: -1/RT ~ Condition + Session + (1|Sub) + (1|Item)

The simple effect of Condition showed that overall, the transformed RTs were significantly longer in the M condition (B = 9.25E-05, $p < .001$) compared to those in the R condition. The same was true for the SA condition (B = 1.27E-04, $p < .001$) compared to the R condition. Overall transformed RTs were also significantly shorter in Session2 compared to Session1 regardless of the condition (B = -9.10E-05, $p < .001$). The overall transformed RTs in the M condition did not differ reliably from those in the SA condition (B = 3.40E-05, $p = .18$) as shown by the planned contrast (see Table 7).

In sum, the RT data did not show semantic priming effect in any of the conditions. The overall RTs were the shortest for processing well-established words (the R condition) and significantly slower for processing direct meanings or semantic associates of the novel words (the M and SA conditions). The overall RTs were also faster after a 48 hour delay compared to the performance immediately after learning the novel words. The lack of semantic priming effect was probably due to considerably long RTs (over 1000 ms for all conditions and over 1500 ms for SA condition in Session1).

Table 7. Model coefficients for RTs in Semantic Relatedness task. Model (Condition + Session) fixed effect coefficient estimates for mean inverse transformed RTs. Reference levels: R condition, Session1.

	Coef. B	Std. Error	df	t-value	p-value	
(Intercept)	-8.99E-04	2.47E-05	42.30	-36.38	7.55E-33	***
ConditionM	9.29E-05	1.57E-05	67.83	5.92	5.80E-07	***
ConditionSA	1.27E-04	6.22E-06	8533.42	20.38	1.69E-89	***
Session	-9.10E-05	5.07E-06	8532.63	-17.93	6.85E-70	***
Contrasts:						
ConditionM vs ConditionSA	3.40E-05	1.59E-05	71.41	2.14	0.18	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. P-value correction: Bonferroni method for 5 tests

Numerically, the observed RTs were longer for unrelated trials than the related trials only in the R condition, whereas this pattern was reversed for both novel word conditions (see Figure 4). See further considerations of the results in the Discussion.

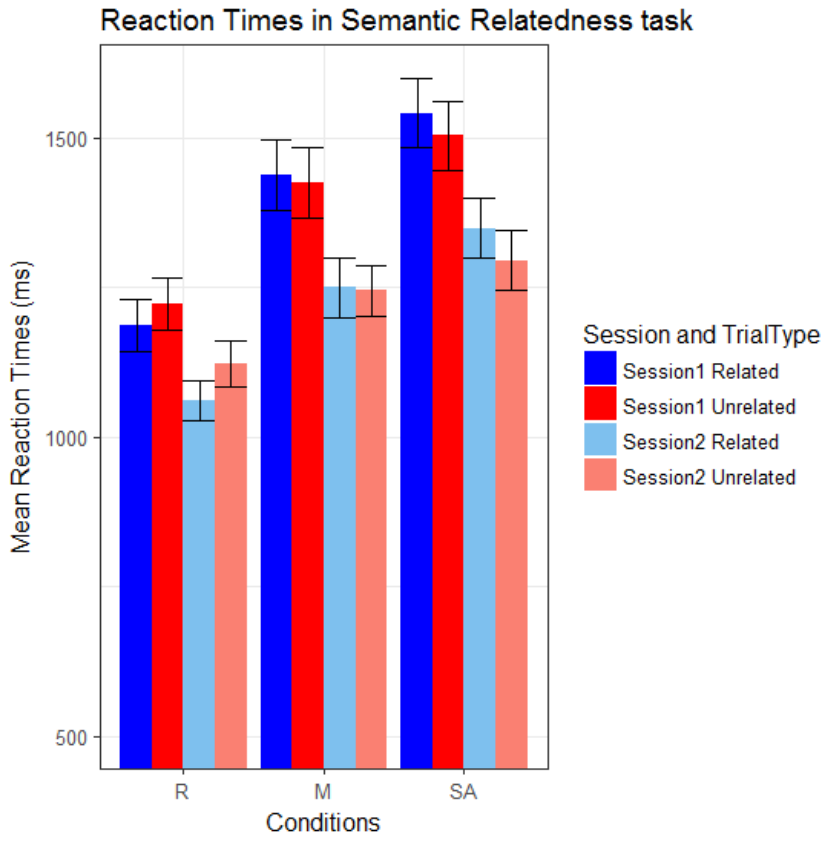


Figure 4. Reaction Times in Semantic Relatedness task. Observed mean reaction times (ms) for Real word (R), Meaning (M) and Semantic Associate (SA) conditions by Session and Trial type, starting at 500 ms. Error bars are standard errors.

3.2.4 Event Related Potentials in Semantic Relatedness task

3.2.4.1 EEG recording and pre-processing

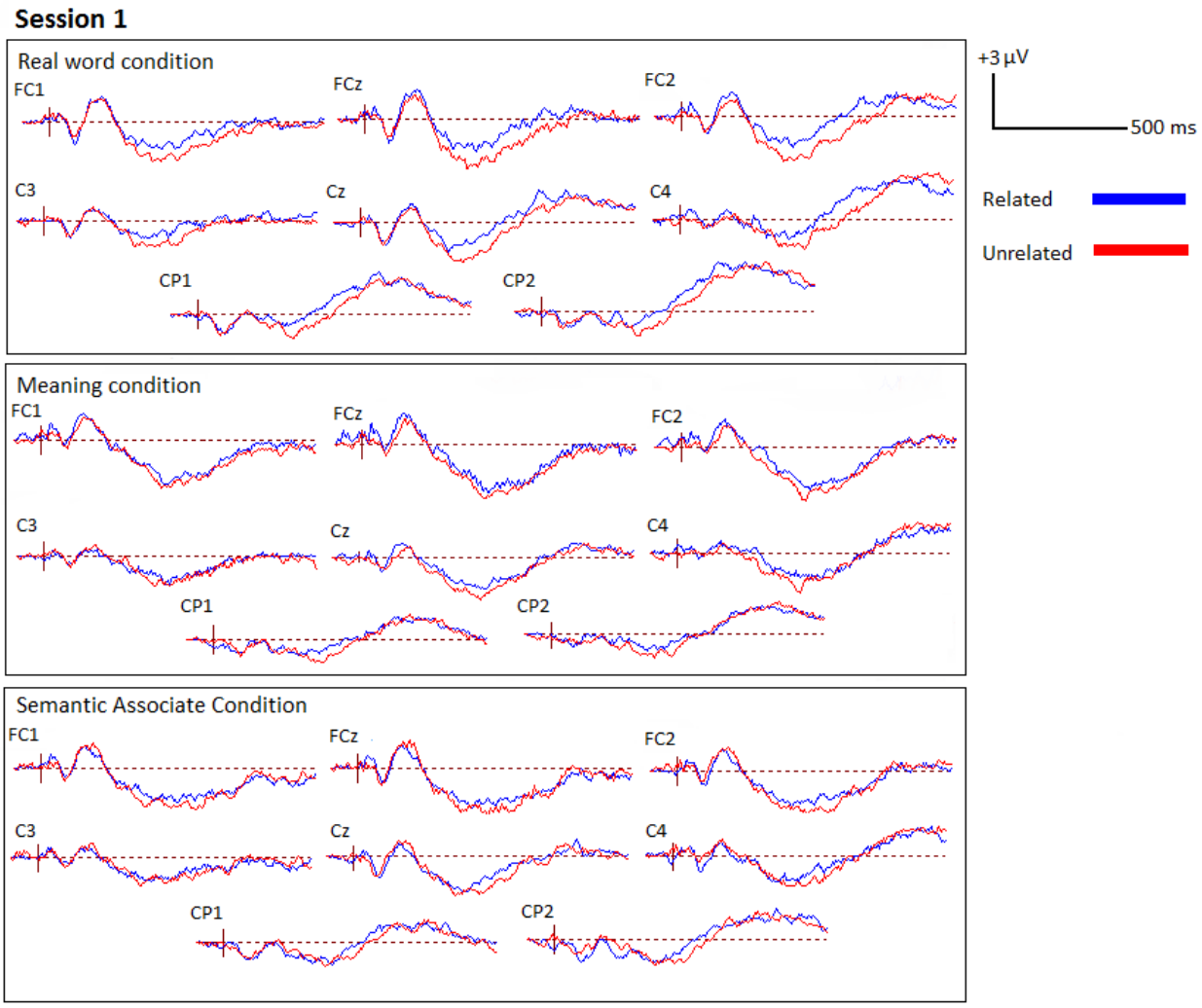
Electrophysiological responses to prime and target words in the Semantic Relatedness task were recorded using BrainVision Recorder (Brain Products GmpH). The recordings were performed in a shielded room with 32 Ag/AgCl elegtrodes, placed according to the extended international 10-20 system. FCz was used as the reference electrode and the EEG data was recorded at 1 kHz sampling rate. Impedances were kept at or below 15 kΩ and a total of 11 electrodes across all subjects and recording sessions exceeded this value. These electrodes were later excluded as bad channels or interpolated during off-line pre-processing of the data. As the maximum level of impedances is higher than the usual 5kΩ, it might reduce the signal-to-noise ratio of the recordings (Kappenman & Luck, 2010). However, active electrodes were used in the recordings, which alleviate the problem of high impedances to certain extent.

EEG data was then pre-processed off-line using BESA ® Research 5.3.7. An automatic, adaptive eye blink correction was applied to the data, after which any remaining blinks were selected manually for correction. Segments of 1000 ms were created relative to the prime or target word onset (100 ms before and 900 ms after the onset of the word). A baseline correction was performed by subtracting the mean amplitude in the 100 ms time window before the word onset. A low pass filter of 40 Hz and high pass filter of 0.1 Hz were applied to the data. The created segments were then scanned for artifacts so that segments with over 140 μV amplitudes, gradients of over 75 μV/sampling interval or signals lower than 0.01 μV were rejected. Bad channels were removed or interpolated. A total of 2% of trials were rejected due to artifacts. The data was re-referenced using an average reference and unfiltered mean amplitudes for time windows 0-300 ms, 300-600 ms and 600-900 ms were exported for further analysis. Two participants were excluded for low quality EEG data in both sessions and one participant was excluded for low quality EEG data in Session 1 (defined as over 20% of trials lost in any one condition or over 6 bad channels). The data for the remaining 26 participants for Session 1 and 27 participants for Session 2 were used in the analyses.

3.2.4.2 Event-related Potentials Analyses

The purpose of the Event-Related Potentials (ERP) analyses was to investigate the lexicalization of the novel words as indexed by the N400 effect: more negative ERP responses to target words preceded by semantically unrelated prime words than to targets preceded by related primes. The N400 effect is expected in centro-parietal electrode locations approximately 300-600 ms after target onset for visual stimuli. In the current analyses with auditory stimuli, the latency range might extend beyond this, which is why 600-900 ms time window was included in the analyses. The ERP data was analyzed with mixed effects linear regression analysis for each time window (0-300 ms, 300-600 ms and 600-900 ms after target onset) separately. Dependent variable in these analyses was the mean Event Related Potential (ERP) amplitude in the region of interest (electrode locations FC1, FC2, FCz, C3, C4, Cz, CP1 and CP2) chosen based on visual inspection of grand averages of the ERP data. This region showed the clearest N400 effect in the data and corresponds to the scalp location for traditionally maximal N400 effect (Kutas & Federmeier, 2011). See Figure 5 for mean ERP waveforms and Figure 6 for summary of the mean ERP amplitudes.

(a)


Session 1
Real word condition
FC1
FCz
FC2
C3
Cz
C4
CP1
CP2
+3 µV
500 ms
Related
Unrelated
Meaning condition
FC1
FCz
FC2
C3
Cz
C4
CP1
CP2
Semantic Associate Condition
FC1
FCz
FC2
C3
Cz
C4
CP1
CP2

(b)

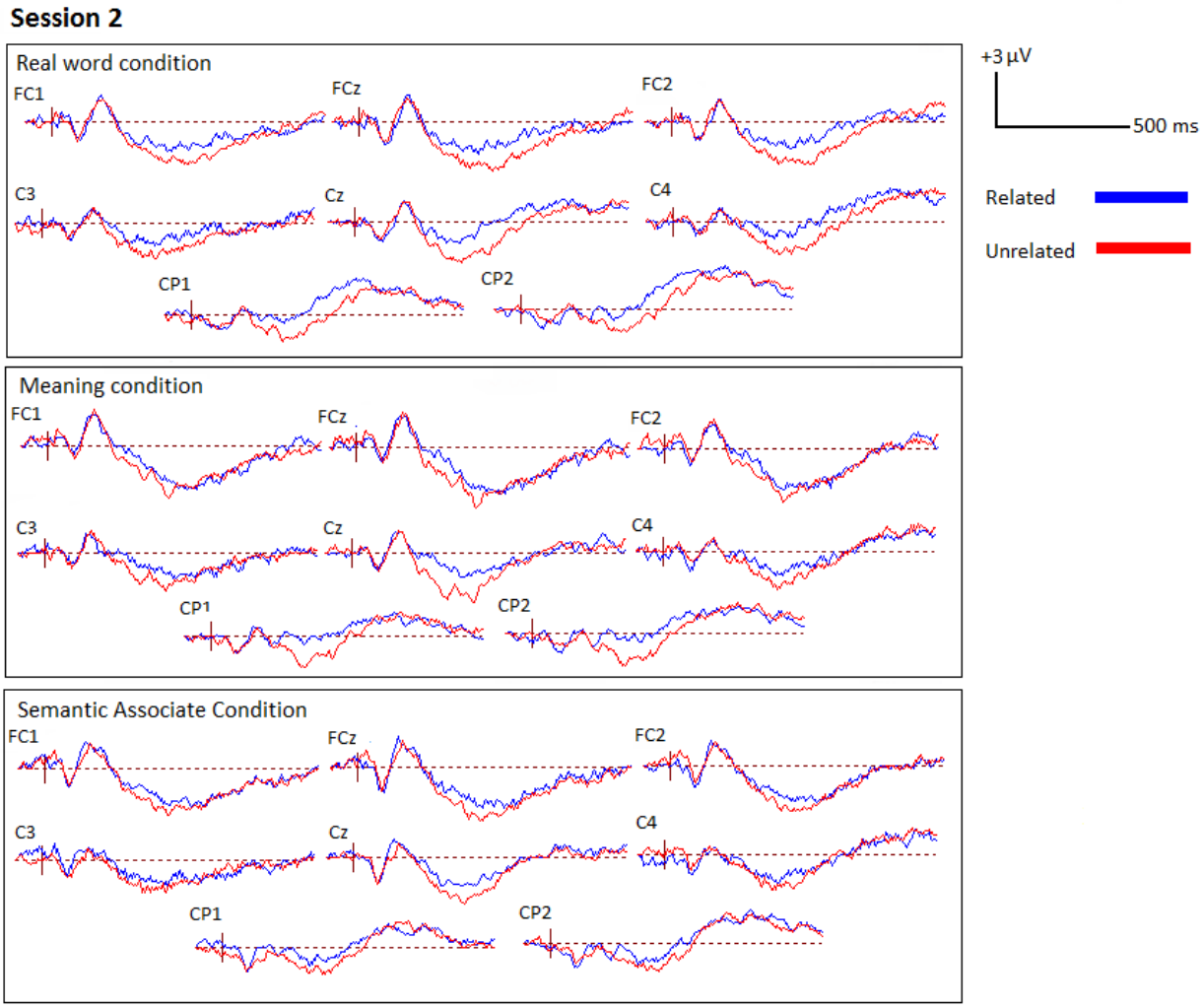


Figure 5. Grand average ERP waveforms. Grand average ERPs to target words in related and unrelated trials in Real, Meaning and Semantic Associate conditions for electrode locations of the region of interest in (a) Session1 and (b) Session2.

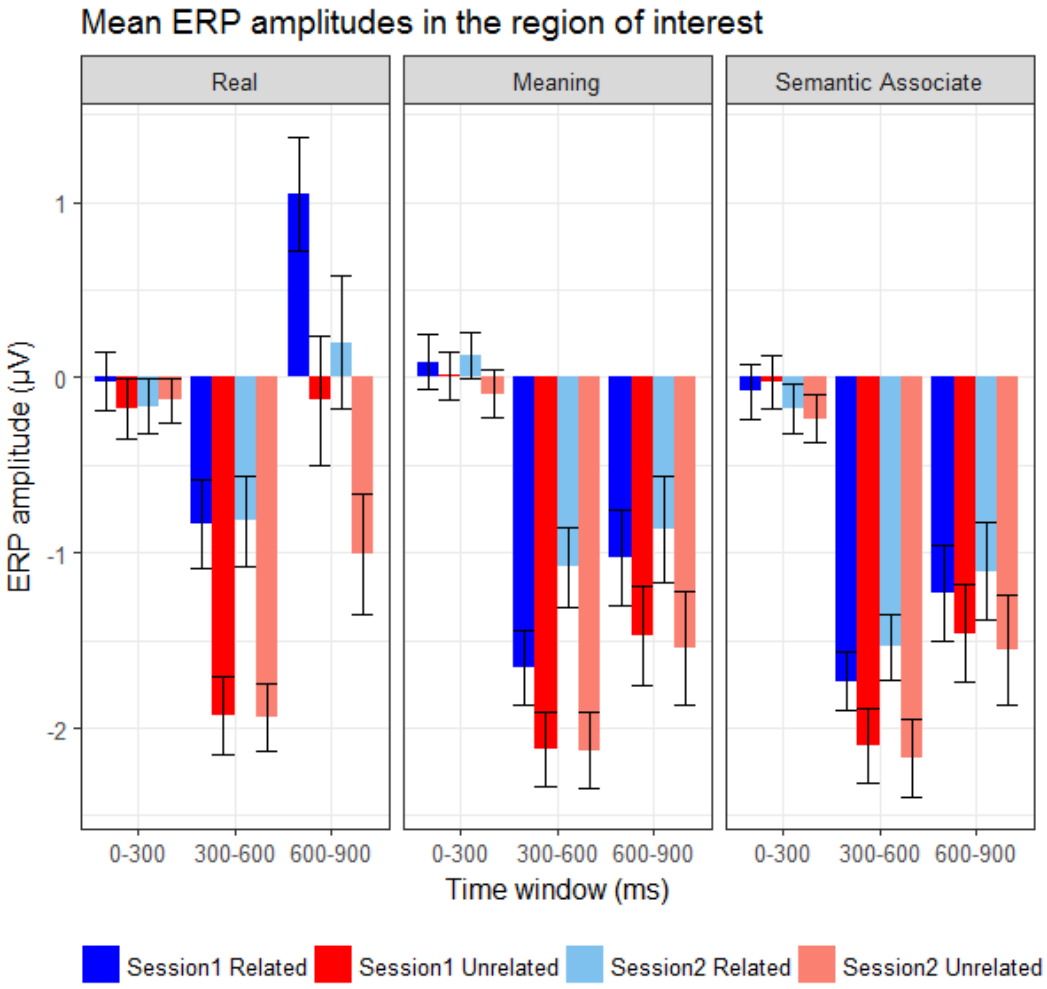


Figure 6. Mean ERP amplitudes in the region of interest. Observed values of ERP amplitudes for Real, Meaning and Semantic Associate conditions by session and trial type. Error bars are standard errors.

The ERP data with a total of 9968 observations was analyzed using LME regression analyses for each time window separately. Treatment coding was used for categorical variables with Real word condition, Session 1 and Related trials as reference levels.

3.2.4.3 *0-300 ms time window*

For visualization, the R formula for the final model (referred to as Model-03) in the 0-300 ms time window was as follows:

Model-03: ERP.0-300 ~ (1|Subject)

The model had no fixed effect predictors and only Subject as a random effect. This suggests that the variation in mean ERP amplitudes was explained only by the individual participants. The model thus shows – as expected – that the mean ERP amplitudes did not differ by Condition, TrialType or Session in this early time window.

3.2.4.4 300-600 ms time window

The R formula for the final model (referred to as Model-36) in the 300-600 ms time window was as follows:

Model-36: ERP.300-600 ~ Condition*TrialType + Condition + TrialType + (1|Subject) + (1|Item)

The fixed effects for Model-36 were Condition, Trial type and an interaction of Condition and Trial type. There was a simple effect of Condition: compared to the ERP responses (mean amplitude) in the R condition, the responses were overall more negative in the M condition ($B = -.54$, $p < .05$) and in the SA condition ($B = -.81$, $p < .001$). There was also a simple effect of Trial type, with the ERP responses to unrelated trials being more negative than responses to related trials ($B = -1.11$, $p < .001$). The simple effects of Condition and Trial type were qualified by an interaction of Condition and Trial type: the difference in ERP responses to the related vs unrelated trials in the M condition was not reliably smaller ($B = .35$, $p = .45$) than that in the R condition, whereas for the SA condition this difference was reliably smaller ($B = .61$, $p < .01$) compared to that in the R condition (see Figure 7). As expected, the difference between ERP responses to the related vs unrelated trials (i.e. the N400 effect) in the R condition was statistically significant ($B=-1.11$, $p < .001$, the simple effect of Trial type), and planned contrasts for the difference between responses to related vs unrelated trials in the M and SA conditions were found statistically significant as well (M condition: $B = -.76$, $p < .001$; SA condition: $B = -.51$, $p < .001$). Finally, a planned contrast comparing the difference in mean ERP amplitudes to related vs unrelated trials in the M condition to that in the SA condition was not statistically reliable ($B = .26$, $p = 1$). See Table 8 for fixed effect coefficients and contrasts for Model-36.

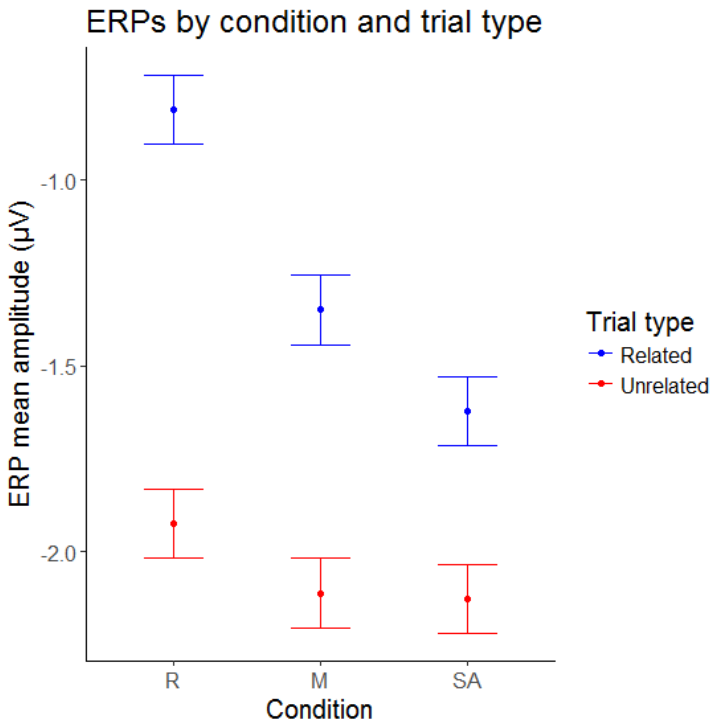


Figure 7. Model-36 estimates of ERPs by condition and trial type. Conditions Real word (R), Meaning (M) and Semantic Associate (SA)) in 300-600 ms time window. Error bars are standard errors.

Table 8. Model-36 coefficients for ERP amplitudes. Model-36 (Condition*TrialType + TrialType + Condition) fixed effects coefficients and planned contrasts in 300-600 ms time window. Reference levels: Condition R, Related trials.

Variable	Coef. B	Std. Error	df	t-value	p-value	
(Intercept)	-0.81	0.18	62.92	-4.48	2.85E-04	***
ConditionM	-0.54	0.17	164.29	-3.25	1.25E-02	*
ConditionSA	-0.81	0.13	9875.82	-6.31	2.69E-09	***
TrialType	-1.11	0.13	9875.30	-8.66	4.86E-17	***
ConditionM*TrialType	0.35	0.18	9875.73	1.93	0.45	
ConditionSA*TrialType	0.61	0.18	9875.42	3.34	7.54E-03	**
Contrasts:						
ConditionM*TrialType vs ConditionSA*TrialType	0.26	0.18	9875.82	1.41	1	
Condition M related vs unrelated trials	-0.76	0.13	9876.22	-5.93	2.88E-08	***
Condition SA related vs unrelated trials	-0.51	0.13	9875.47	-3.93	7.59E-04	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. P-value correction: Bonferroni method for 9 tests

Although there were no effects of Session in the final model, the data from Session 1 and Session 2 were analyzed separately to investigate the emergence of the N400 effect in more detail. Model-36 was fitted to the data from Session1 and Session2 separately and this model will be referred to as Model-bySession when data from only one testing session was used. These additional analyses revealed that in Session 1, the responses to related vs unrelated trials did not differ reliably in M ($B = -.47$, $p = .09$) or in SA ($B = -.38$, $p = .36$) condition, although the trial type difference in these novel word conditions was not reliably smaller than that in the R condition (M condition vs R condition trial type difference: $B = .62$, $p = .18$; SA condition vs R condition trial type difference: $B = .72$, $p = .06$). The Trial type difference was only reliable in the R condition ($B = -1.09$, $p < .001$). In Session 2, in addition to the responses to related vs unrelated trials being reliably different in the R condition ($B = -1.12$, $p < .001$), they were also reliably different in both M ($B = -1.04$, $p < .001$) and SA condition ($B = -0.63$, $p < .01$). The trial type difference in the novel word conditions did not differ from that in the R condition (M condition vs R condition trial type difference: $B = .08$, $p = 1$; SA condition vs R condition trial type difference: $B = .49$, $p = .45$). See Tables 6 and 7 for fixed effects coefficients and contrasts for Model-bySession for Session1 and Session2. See Appendix C for more detailed considerations of the separate analyses for each testing session.

A likely reason for statistically non-significant effects of testing session in the analysis run with the data from both sessions is lack of statistical power. Even though the Model-36 is the best fit for the data, it would be premature to conclude that the learning effects were present in both sessions. The observed values (See Figure 5 and 6, p. 51-53) showed modest N400 effects in the novel word conditions in Session 1 and considerably larger N400 effects in these conditions in Session 2. Given this clear numerical trend in the observed data and the information provided by the separate analyses for the testing sessions, it is concluded that the semantic learning of the novel words as indexed by reliable N400 effect was not found in the first testing session, but a reliable N400 effect was found in the second session, where the magnitude of the N400 effect for novel words was comparable to that for real words. These results demonstrate that the novel words were learnt well enough to show signs of lexicalization, but this learning effect only emerged after a 48 hour delay.

Table 9. Model-bySession coefficients for Session 1 ERP amplitudes in 300-600 ms time window. Model-bySession (Condition*TrialType + Condition + TrialType) fixed effects coefficients and planned contrasts. Reference levels: Condition R, Related trials.

Variable	Coef. B	Std. Error	df	t-value	p-value	
(Intercept)	-0.83	0.22	84.62	-3.87	1.95E-03	**
ConditionM	-0.82	0.23	196.92	-3.61	3.54E-03	**
ConditionSA	-0.90	0.19	4826.49	-4.83	1.24E-05	***
TrialType	-1.09	0.19	4826.39	-5.88	4.02E-08	***
ConditionM*TrialType	0.62	0.26	4826.73	2.37	0.18	
ConditionSA*TrialType	0.72	0.26	4826.21	2.72	0.06	.
Contrasts:						
ConditionM*TrialType vs ConditionSA*TrialType	0.09	0.26	4826.66	0.35	1	
Condition M related vs unrelated trials	-0.47	0.19	4827.14	-2.52	0.09	.
Condition SA related vs unrelated trials	-0.38	0.19	4826.24	-2.03	0.36	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. P-value correction: Bonferroni method for 9 tests

Table 10. Model-bySession coefficients for Session 2 ERP amplitudes in 300-600 ms time window. Model-bySession (Condition*TrialType + Condition + TrialType) fixed effects coefficients and planned contrasts. Reference levels: Condition R, Related trials.

Variable	Coef. B	Std. Error	df	t-value	p-value	
(Intercept)	-0.82	0.22	58.50	-3.77	3.42E-03	**
ConditionM	-0.27	0.19	279.05	-1.38	1	
ConditionSA	-0.72	0.18	4958.65	-4.09	4.01E-04	***
TrialType	-1.12	0.18	4957.62	-6.41	1.47E-09	***
ConditionM*TrialType	0.08	0.25	4958.12	0.33	1	
ConditionSA*TrialType	0.49	0.25	4958.15	1.97	0.45	
Contrasts:						
ConditionM*TrialType vs ConditionSA*TrialType	0.41	0.25	4958.14	1.65	0.9	
Condition M related vs unrelated trials	-1.04	0.18	4958.62	-5.95	2.57E-08	***
Condition SA related vs unrelated trials	-0.63	0.18	4957.68	-3.61	2.77E-03	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. P-value correction: Bonferroni method for 9 tests

3.2.4.5 600-900 ms time window

The R formula for the final model (referred to as Model-69) in 600-900 ms time window was as follows:

Model-69: ERP.600-900 ~ Condition*Session + Condition*TrialType + TrialType + Condition + Session + (1|Subject) + (1|Item)

There was a simple effect of Condition: the overall ERP responses in both of the novel word conditions were more negative than responses in the R condition (M condition: B = -2.02, p < .001; SA condition: B = -2.23, p < .001). There was also a simple effect of Session: the ERP responses in the R condition were overall more negative in Session 2 compared to Session 1 (B = -.90, p < .001). Additionally, there was a simple effect of Trial type, where ERP responses to unrelated trials were more negative than responses to related trials (B = -1.2, p < .001). The simple effects of Condition and Session were qualified by an interaction: compared to the difference in the overall ERP amplitudes between Session 1 and Session 2 in the R condition, the difference was smaller in the M condition (B = .91, p < .001) and in the SA condition (B = .88, p < .001). This difference in the overall ERP amplitudes between sessions was comparable in the two novel word conditions, as shown by a planned contrast (B = -.03, p = 1). Finally, the simple effect of Trial type was qualified by an interaction of Condition and Trial type, where the difference in ERP responses to related vs unrelated trials was not reliably smaller in the M condition (B = .63, p < .05), but significantly smaller in the SA condition (B = .85, p < .001) compared to that in the R condition (see Figure 8). The planned contrast comparing the difference in responses to related vs unrelated trials in the M condition to that in the SA condition was not significant (B = 0.22, p = 1). However, the planned contrasts for the difference between ERP mean amplitudes to related vs unrelated trials in the two novel word conditions (M and SA) showed that this difference was statistically significant in the M condition (B = -.57, p < .001), but not in the SA condition (B = -.35, p = .18). See table 11 for fixed effect coefficients and contrasts for Model-69.

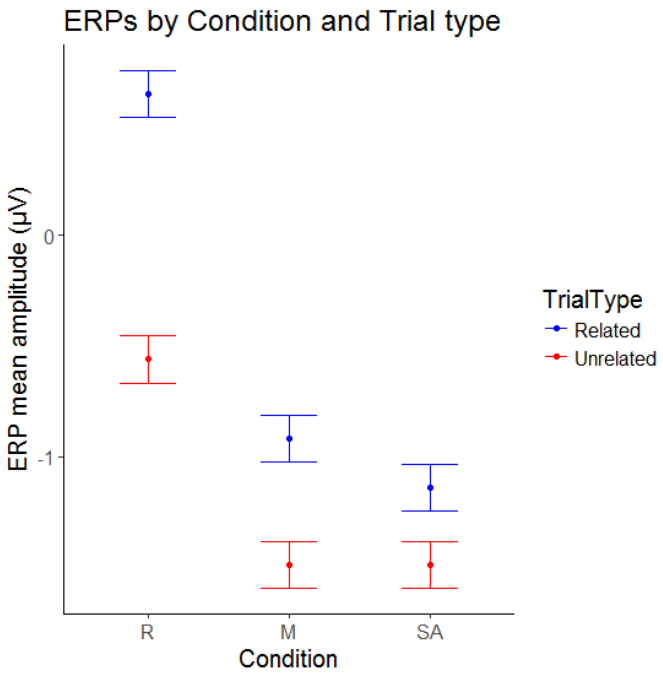


Figure 8. Model-69 estimates of ERPs by condition and trial type. Conditions Real word (R), Meaning (M) and Semantic Associate (SA)) in 600-900 ms time window. Error bars are standard errors.

Table 11. Model-69 coefficients for ERP amplitudes. Model-69 (Condition*Session + Condition*TrialType + TrialType + Condition + Session) fixed effects coefficients and planned contrasts in 600-900 ms time window. Reference levels: Condition R, Related trials, Session 1.

Variable	Coef. B	Std. Error	df	t-value	p-value	
(Intercept)	1.10	0.28	46.23	3.98	3.15E-03	**
ConditionM	-2.02	0.20	335.89	-9.88	2.95E-19	***
ConditionSA	-2.23	0.18	9872.45	-12.58	6.50E-35	***
Session	-0.90	0.14	9878.49	-6.25	5.45E-09	***
TrialType	-1.20	0.14	9872.28	-8.34	1.09E-15	***
ConditionM*Session	0.91	0.20	9872.35	4.49	9.54E-05	***
ConditionSA*Session	0.88	0.20	9871.97	4.33	1.95E-04	***
ConditionM*TrialType	0.63	0.20	9872.76	3.10	2.52E-02	*
ConditionSA*TrialType	0.85	0.20	9872.44	4.18	3.78E-04	***
Contrasts:						
ConditionM*Session vs ConditionSA*Session	-0.03	0.20	9872.39	0.15	1	
ConditionM*TrialType vs ConditionSA*TrialType	0.22	0.20	9872.86	1.08	1	
ConditionM related vs unrelated trials	-0.57	0.14	9873.26	-3.96	9.98E-04	***
ConditionSA related vs unrelated trials	-0.35	0.14	9872.49	-2.42	0.26	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. P-value correction: Bonferroni method for 13 tests

Again, to investigate the emergence of the N400 effect in more detail, separate analyses were run for Session 1 and Session 2. The best fit for data from both sessions was the same model (Model-bySession) used in the 300-600 ms time window (See Table C3, Appendix C). These analyses revealed that in Session 1, the N400 effect (difference in responses to related vs unrelated trials) was not reliable in either of the novel word conditions (M condition: B = -.46, p = .25; SA condition: B = -.24, p = 1) as shown by planned contrasts, but it was statistically significant in the R condition (B = -1.18, p < .001). In Session 2, the N400 effect was reliable in M (B = -.67, p < .01) and R condition (B = -1.21, p < .001), and the magnitude of the N400 effect didn't differ reliably between these two conditions (B = .54, p = .44) as shown by the ConditionM*TrialType interaction. The N400 effect was not found in the SA condition (B = -.46, p = .17) in Session 2. See Table 12 and 13 for fixed effect coefficients and contrasts for Model-bySession in Session 1 and Session 2.

Table 12. Model-bySession coefficients for Session 1 ERP amplitudes in 600-900 ms time window. Model-bySession (Condition*TrialType + Condition + TrialType) fixed effects coefficients and planned contrasts. Reference levels: Condition R, Related trials.

Variable	Coef. B	Std. Error	df	t-value	p-value	
(Intercept)	1.05	0.31	49.50	3.43	0.01	*
ConditionM	-2.07	0.25	205.69	-8.21	2.17E-13	***
ConditionSA	-2.27	0.21	4827.62	-10.91	1.97E-26	***
TrialType	-1.18	0.21	4827.52	-5.67	1.36E-07	***
ConditionM*TrialType	0.72	0.29	4827.85	2.45	0.13	
ConditionSA*TrialType	0.94	0.29	4827.35	3.19	0.01	*
Contrasts:						
ConditionM*TrialType vs ConditionSA*TrialType	0.22	0.29	4827.78	0.74	1	
Condition M related vs unrelated trials	-0.46	0.21	4828.23	-2.19	0.25	
Condition SA related vs unrelated trials	-0.24	0.21	4827.37	-1.15	1	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. P-value correction: Bonferroni method for 9 tests

Table 13. Model-bySession coefficients for Session 2 ERP amplitudes in 600-900 ms time window. Model-bySession (Condition*TrialType + Condition + TrialType) fixed effects coefficients and planned contrasts. Reference levels: Condition R, Related trials.

Variable	Coef. B	Std. Error	df	t-value	p-value	
(Intercept)	0.20	0.32	37.81	0.65	1	
ConditionM	-1.07	0.20	398.65	-5.31	1.65E-06	***
ConditionSA	-1.30	0.19	4960.74	-6.67	2.52E-10	***
TrialType	-1.21	0.19	4959.29	-6.24	4.21E-09	***
ConditionM*TrialType	0.54	0.28	4959.55	1.97	0.44	
ConditionSA*TrialType	0.76	0.28	4960.05	2.74	0.05	.
Contrasts:						
ConditionM*TrialType vs ConditionSA*TrialType	0.21	0.28	4959.57	0.78	1	
Condition M related vs unrelated trials	-0.67	0.19	4959.79	-3.46	4.85E-03	**
Condition SA related vs unrelated trials	-0.46	0.19	4959.36	-2.36	0.17	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. P-value correction: Bonferroni method for 9 tests

In sum, the N400 effect found in the previous time window (300-600 ms) was still present in the late time window (600-900 ms) when processing real words (the R condition) in both testing sessions. The N400 effect was also found when processing direct meanings of the novel words (the M condition) in Session 2 only. However, when novel word processing required connecting the novel word meaning to its semantic field (the SA condition), the N400 effect was no longer statistically reliable in either of the testing sessions.

4. Discussion

The current study explored lexical representations of novel word forms and meanings acquired via aurally presented sentence contexts and the effects of language proficiency in this type of novel word acquisition. In order to elucidate whether this learning mode would result in episodic representations of the novel words (e.g. successful recognition of the novel words) and whether these representations become more integrated into the semantic memory (e.g. effects of the newly learnt words on the processing of known words), the novel word acquisition was tested at two different testing sessions utilizing explicit and implicit measures of word knowledge. The findings from the analyses of word form knowledge will be discussed first, followed by findings from the analyses of word meaning knowledge. The L1 and L2 proficiency effects in the above mentioned dimensions of novel word acquisition will be discussed subsequently. Finally, limitations and suggestions for further research will be discussed, followed by conclusions from the current study.

4.1 Word form acquisition

The explicit knowledge of the novel word forms was tested with 2AFC task. The recognition accuracy of the novel word forms was reliably above chance when tested almost immediately after learning the novel words and after a 48 hour delay. The response accuracy also increased reliably across testing sessions.

The implicit knowledge of the newly learnt word forms was tested with Pause Detection task. In this task, lexicalization of the novel word forms is characterized by slower pause detection responses to known words that resemble recently learnt novel words compared to control words after a delay, but not immediately after learning the novel words. The only finding in the current study was that items with pauses were responded to slower than items without pauses. Thus, the analysis did not reveal a lexicalization effect in either of the testing sessions, suggesting that the representations of the newly learnt word forms were not lexicalized, (e.g. recognition of a well-established *cathedral* was not hindered by earlier exposure to a novel word *cathedruke*). A likely reason for this is the task difficulty: participants were exposed to

each of the 32 novel word forms only 15 times by the end of Session 1 (including contextual learning task, 2AFC task, Semantic Relatedness task and the Restudy task). This is considerably less exposure than what is typically reported in the literature (e.g. Gaskell & Dumay, 2003; Dumay & Gaskell, 2007). Additionally, participants had to divide their cognitive resources between learning the word forms and linking them to meanings, whereas most Pause Detection experiments have only focused on word form learning (but cf. Henderson, Devine, Weighall & Gaskell, 2015; Dumay, Gaskell & Feng, 2004). However, a numeric trend for a lexicalization effect was found in the current RT data. Interestingly, Dumay, Gaskell & Feng (2004, Experiment 1) found differing developmental trajectories of lexicalization effect for word forms that were learnt based on phonological exposure only and word forms that were learnt via semantic exposure. Utilizing Lexical Decision task, they found lexicalization effect for the phonologically learnt items already 24 hours after the initial exposure, whereas this effect was found for the semantically learnt items only a week after the initial exposure. These results would suggest that lexicalization of novel word forms learnt together with their meanings takes longer than word forms learnt in isolation. However, Henderson et al. 2015 showed lexicalization effect for contextually learnt items only 24 hours after the initial exposure, when 12 items were learnt and there were 12 exposures to each item. Therefore, it is possible that the 48 hour delay in the current study was not sufficient for the novel word forms to be fully lexicalized, given the amount of items to be learnt and the amount of exposure for each item.

Taken together, although the novel word forms were learnt, they had not been fully lexicalized, potentially because the learning and testing procedure utilized was not optimal for a strong enough lexicalization of the novel word forms to take place or to be detected. This pattern of results suggests a dissociation in the emergence of explicit and implicit word form knowledge, as the between sessions improvement in explicit recognition accuracy observed in the 2AFC task was not accompanied with implicit word form knowledge in the Pause Detection task. This dissociation is in line with the complementary systems account of word learning and has been observed in several other studies, where lexicalization has been found after a delay. Importantly, the later emergence of lexicalization effects compared to signs of explicit word form knowledge have been found regardless of whether the novel words were learnt in

isolation or linked with their meaning (e.g. Gaskell et al. 2003; Dumay et al. 2004; Henderson et al. 2015).

4.2 Word meaning Acquisition

Explicit knowledge of the novel word meanings was tested via the Semantic Relatedness task, where participants made relatedness judgments of word pairs that were either two known words (R condition), a novel word and its meaning (M condition) or a novel word and its semantic associate (SA condition). Response Accuracies in this task showed clear above chance level performance in all three conditions both immediately after the word learning task and after a 48 hour delay. The response accuracies also increased reliably over time for the novel word conditions (M and SA conditions), where judgments for semantic relatedness were made between a novel word and its meaning or semantic associate. Crucially, this learning as indexed by increased recognition accuracy was only seen in the novel word conditions, but not in the control condition (R condition), where the semantic judgments were made between two known words. This further confirms that the gains in recognition accuracy over time were not merely due to task familiarity or repetition of the test items in Session 1 and Session 2. Additional support for this finding comes from the multiple regression analyses run for the data in Semantic Relatedness task: response accuracy in the first session predicted the response accuracy in the second session for the novel word conditions, but not reliably so in the real word condition. It is concluded that contextual novel word learning task in the current study produced reliable explicit knowledge of the novel word meanings and this knowledge supported successful semantic judgments even in cases that extended beyond the direct meanings of the novel words.

Implicit acquisition of novel word meanings was also tested with Semantic Relatedness task. The indices of lexicalization of the novel words used were the semantic priming effect and the N400 effect: a pattern of reaction times (RT) and the mean amplitudes of Event-Related Potentials (ERP) where responses to words following a semantically related word would be faster (RTs) or less negative (ERPs) than responses to words following a semantically unrelated word. The semantic relatedness effect was not observed in the RT data, where the overall

responses were faster after a 48 hour delay than immediately after learning the novel words. Additionally, the RTs in both of the novel word conditions (M and SA conditions) were reliably slower than those in the real word (R) condition. There was a numeric trend for semantic relatedness effect in the R condition only, whereas both novel word conditions showed numerically larger RTs for related trials compared to unrelated trials. This lack of priming effect is likely due to considerably slow RTs and a considerably large number of excluded trials in the novel word conditions. The slow RTs might reflect the task difficulty or focus on accuracy over speed, although why semantic judgments between two known words (the R condition) would be this effortful, is unknown. Interestingly, similar pattern of RT data was observed by Frishkoff et al. (2010) in their contextual word learning experiment, where known words, trained novel words and untrained novel words all elicited faster RTs for unrelated trials compared to related trials. Similarly, Mestres-Misse et al. (2007) reported faster RTs for unrelated trials for both known words and novel words immediately after the training phase. In both of these studies, as well as in the current study, the RT data was collected simultaneously with ERP data collection, which might pose too many competing instructions for the participants: to respond as quickly and accurately as possible, while not blinking during certain parts of the stimulus presentation. However, this type of procedure is used widely, whereas unsuccessful discoveries of priming effects in RT data are reported less often. It is concluded that the RTs in the Semantic Relatedness task in the current study were exceptionally slow and no semantic relatedness effects were found.

The ERP data showed the N400 effect in the expected 300-600 and 600-900 ms time windows after the target onset. Whereas known words (the R condition) elicited the N400 effect in both testing sessions, the effect was not found immediately in the novel word conditions (M and SA conditions), but only after a 48 hour delay. The N400 effect was found for judgments made between two known words (R condition) and between a novel word and its meaning (M condition) in the expected time windows (300-600 ms and 600-900 ms). For judgments between a novel word and its semantic associate (SA condition) the N400 effect was found only in the 300-600 ms time window, which is potentially indicative of a weaker learning effect in this condition. These results are considered in the light of previous studies in contextual novel

word acquisition in visual domain, both for immediate (Mestres-Misse et al., 2007) and delayed test of meaning acquisition (Elgort et al. 2015; Frishkoff et al., 2010). As the studies by Mestres-Misse et al. (2007) and Frishkoff et al. (2010) used the same semantic relatedness task and novel word condition (the Meaning condition in the current study) as the current study, a number of comparisons can be made. Firstly, the latency of the N400 effect for novel words in these studies was reported as 300-500 and 500-700 ms for immediate test (Mester-Misse et al., 2007) and 350-450 ms for a test 48 hours after the initial learning (Frishkoff et al., 2010). By contrast, in the current study the N400 effect was found in 300-600 and 600-900 ms time windows. Although the choice of time windows made by experimenters naturally affects the reported N400 effects found in each study, the difference in the N400 latency is also not surprising as the latency of auditory N400 effects are known to last longer than their visual counterparts (Holocomb & Neville, 1990). Secondly, in the study by Mestres-Misse et al. (2007), the semantic relatedness judgments were made immediately after blocks of sentence contexts where the novel word meanings were learnt (blocks of 32 sentence triplets where 8 of them were the Meaning condition comparable to the current study). In the current study, by contrast, the semantic relatedness task was administered after several distraction tasks, and the N400 effect was not found in the first testing session. This suggests that the N400 effect for newly learnt words before off-line consolidation might not be resilient enough against distraction, even though the effect can be found immediately after learning the novel words. Thirdly, the previous studies demonstrate a connection between a novel word form and its meaning only. The current study provides a crucial addition to this, demonstrating a connection between a novel word form and its semantic associate. As argued in the Introduction (The Current Study section), this connection provides stronger evidence for lexicalization of the novel words. The observed priming effect between the novel word and its semantic associate is interpreted as the novel words' ability to activate a wider set of items in its semantic field (either directly or via mediation of the meaning of the novel word), just like well-known real words do. This priming effect also demonstrates that the learnt word forms and meanings are not a separate association from the mental lexicon. Note that the prime-target pairs in the Semantic Associate condition were the same ones in Session1 as in Session2, which weakens

the argument slightly, as the participants might have learnt isolated associations between these particular items (e.g. *cathedruke-weave* word form-semantic associate pair might have been learnt in addition to the original *cathedruke-basket* word form -meaning pair). However, this interpretation is unlikely, given that each pair of word form – semantic associate was presented only once in the related condition during the semantic relatedness task in Session1. Instead, it is more likely that the total of 7 encounters of each novel word in semantically supportive contexts resulted in a strong enough link between the novel word and its meaning to allow priming effects between the novel word and its semantic associate as well. Another counter argument for the interpretation of the semantic relatedness effect in the Semantic Associate condition is that the connection between the novel word form and its semantic associate might reflect a vague link between the novel word and its meaning and that's why both the exact meaning of the novel word as well as the semantic associate of the novel word would produce a priming effect. However, as described in the Materials section for Sentence contexts, only 4 items of the total of 32 (12.5%) could plausibly have supported learning of the semantic associate instead of the actual meaning of the novel words. Unfortunately the testing sessions were too long to add an explicit recall task for the novel word meanings that would have allowed detailed investigation of the actual meanings the participants derived from the sentence contexts. Even with this limitation, the vast majority of the learnt items can be assumed to be the actual meanings of the novel words. This conjecture is supported by the pre-test of the sentence contexts (see Sentence Contexts in Materials), where the independent 26 participants consistently filled in the actual meanings of the novel words (the cloze probability for the set of 6 sentences for each item was on average 91.5%). It is concluded that contextually learnt spoken words in the current study resulted in lexicalized semantic items that behave like real words. To our knowledge, this is the first study demonstrating that contextually learnt, spoken novel items affect the activation of their semantic field, not only the specific meanings they are linked to.

4.3 Language proficiency effects

Language proficiency effects found in the current study in tests of explicit word form and word meaning knowledge will be discussed first, followed by a general discussion of the language proficiency effects in the current study.

The multiple regression analyses revealed that word form recognition accuracy in the 2AFC task was associated with language proficiency: higher L1 proficiency was linked to higher immediate and delayed recognition accuracy, whereas higher L2 proficiency was linked to higher gains in recognition accuracy between sessions. Similar type of observation was made by Henderson et al. 2015, where children with larger expressive vocabulary showed larger improvement over time in both explicit and implicit measures of word form acquisition when the novel words were learnt from a context. Their study demonstrated a within language benefit (L1 vocabulary was linked to larger gains in L1 contextual word learning). Henderson et al. interpret this finding as the "Matthew effect" of word acquisition – larger vocabularies support further acquisition of novel words. This interpretation might fit the current results of L2 effect as well: if a shared storage for L1 and L2 vocabulary is assumed, this larger resource of phonological forms (compared to only L1 vocabulary) might indeed facilitate the addition and especially consolidation of new items. Another possibility is that the relationship between a vocabulary size and efficiency in novel word acquisition is not based on the richness of the phonological network itself, but rather on a more efficient learning mechanism that has been trained through L2 vocabulary acquisition. People with larger L2 vocabularies would have had more practice in this particular type of learning. These two explanations are not mutually exclusive, and can further be combined with a third option, where individual aptitude to novel word acquisition facilitates development of vocabulary size. However, the role of L1 proficiency in novel word form learning was more influential in absolute performance level in novel word form acquisition, as L1 was associated to both immediate and delayed recognition accuracy. The same explanations suggested for L2 effect above largely apply to L1 effects as well: where the phonological network of a highly proficient L1 speaker might not have the same richness (i.e. variability of phonological sequences) that comes from high level of L2 knowledge, it would

still have a wider range of native phonological sequences than in a phonological network of a less proficient L1 speaker. This larger phonological repertoire in long-term memory might facilitate acquisition of new phonological sequences by easing the demands on the STM during encoding. Regardless of the underlying mechanisms for the association between language proficiency and word form acquisition, it is concluded that higher immediate and delayed recognition accuracy of novel word forms learnt from aurally presented contexts is associated to higher L1 proficiency.

Analyses of language proficiency effects in explicit word meaning knowledge revealed that there were no reliable predictors for the immediate or delayed performance in the R condition, where semantic relatedness judgments were made between two known words. A possible reason for these results is the near ceiling effect for performance in this condition. L1 proficiency and STM capacity explained the response accuracy for the novel words immediately after learning them as well as after a 48 hour delay in the SA condition, but reliably only after a delay for the M condition. Considering these predictors in word learning, the role of STM capacity in novel word acquisition is well known, especially in learning of novel word forms (e.g. Baddeley, Gathercole & Papagno, 1998; Gupta, 2003). Admittedly, some level of word form knowledge is necessary for meaning acquisition to take place. In the semantic relatedness task in the current study, the word form was given, but needed to be recognized before judgments about its meaning could be made. As such, the role of STM capacity is understandable. The effects of L1 proficiency in the performance of relatedness judgments for novel words could be understood, as suggested above for novel word form acquisition, via the Matthew effect of novel word acquisition: a richer semantic network provides more opportunities to attach new items to.

In sum, higher explicit learning success of novel word forms was associated with higher L1 proficiency both immediately and after a delay, whereas higher L2 proficiency was linked to gains over time in word form recognition. Successful explicit learning of novel word meanings was linked to higher L1 proficiency and higher STM capacity, especially 48 hours after learning the novel words.

In the current study, the main interest in language proficiency effects was in the potential cross-language facilitatory effects of L2 on L1 word acquisition, with a secondary interest in within language benefits (that is, L1 proficiency benefitting novel word learning in L1). As discussed above, higher L2 proficiency was only linked to higher gains in word form recognition accuracy between testing sessions. Contrary to this limited finding, a bilingual advantage in word learning has been observed in previous studies, especially in highly proficient bilinguals (e.g. Kaushanskaya & Marian, 2009; Kaushanskaya, 2012) and even in intermediately proficient late bilinguals (van Hell & Mahn, 1997; Nair, Biedermann & Nickels, 2015). In these studies word learning was measured as recognition and recall of the correct meaning when the word form was provided or vice versa, thus probing both word form and word meaning acquisition. No such L2 effects in novel word meaning acquisition were found in the current study. Considering the suggested explanation for bilingual benefit in word learning (see Introduction, p. 15), namely, a richer semantic network in bilinguals, it is possible that the gains from L2 vocabulary in semantic richness of the mental lexicon only happen with high or near native levels of proficiency in L2, when the meanings of lexical items have become more nuanced. The overall L2 proficiency level in the current study was relatively low and the number of participants with high L2 proficiency was small: only 6 participants scored higher than 50% in the L2 vocabulary test used, whereas native speakers score on average 90% in this test (Izura, Cuetos & Brysbaert, 2014). As such, the language proficiency effects in semantic processing of the novel words might not be detectable because the L2 proficiency in the current sample was not sufficiently high. However, the L1 proficiency across the participants was more varied than expected and the accuracy of semantic judgments was found to vary as a function of this proficiency. As stated above in the discussion of the results from the multiple regression analyses, the mechanisms through which L1 or L2 proficiency might facilitate novel word acquisition can be assumed to be shared or highly similar. As suggested before, the plausible differences between facilitatory effects of L1 and L2 in word learning might be the type of phonological and semantic richness in the mental lexicon that can only be achieved by knowledge of a second language (e.g. language specific connotations in meanings and variability in phonological repertoire from two language systems), not by high proficiency levels

of one language alone. This idea would assume shared storage or interconnectedness of lexical representations for L1 and L2 items, an idea supported by findings of between-language lexical competition (Marian & Spivey, 2003). Nevertheless, higher L1 proficiency alone is enough to support more efficient novel word acquisition. Literature on vocabulary acquisition has ample reports of within language connection between high language skills and efficiency in visual modality (e.g. Jenkins, Stein, & Wysocki, 1984; Bolger et al., 2008; Perfetti et al. 2005; For L2: Pulido, 2003; Elgort et al. 2015). The current study adds to this body of literature with the found association between higher L1 vocabulary size and higher explicit novel word form and meaning acquisition in auditory domain. However, as the L1 and L2 proficiency levels in the current sample had a trend of correlation ($r = .34$, $p=.07$), the view of general individual aptitude in language acquisition remains a conceivable explanation for the association between word learning and language proficiency. As the observed relationship between L1 proficiency and word acquisition is correlational in nature, it does not allow causal inferences in favor of the Matthew effect or the individual aptitude account. Nevertheless, if the findings in the current study are approached from the position of the Matthew effect in novel word acquisition, higher L1 proficiency was found to facilitate novel word and meaning acquisition from context. A tentative conclusion in regards to L2 proficiency from this position is that the facilitating effects on novel word acquisition seem to become detectable only with relatively high levels of L2 proficiency. However, the gains from L2 proficiency in explicit knowledge of novel word forms might be detectable earlier than in other areas of novel word acquisition.

4.4 Limitations and Further Research

By far the most influential limitation in the current study was the low number of participants with high L2 proficiency. The effects of L2 proficiency were expected to be small, which is why detecting any such effects would have required a larger sample size and especially larger number of highly proficient L2 speakers. Pre-screening of participants could have been a way to ensure a more balanced sample of L2 speakers and with more comparable levels of L1 between low and high proficiency L2 speakers. Another considerable challenge in the current study was the task difficulty: although ceiling effects were successfully avoided in every task (except for

real word condition where ceiling effects were expected), the number of exposures for the novel words turned out to be too low especially for detection of lexicalization of the novel word forms. Additionally, the design of the current study does not allow strong conclusions about whether language proficiency facilitates novel word acquisition or whether individual aptitude for word learning results in higher language proficiency, as the found relationship was correlational. This is a limitation in most studies investigating bilingual benefit in word learning. Instead of cross-sectional designs, this topic should be investigated with a longitudinal design in order to inspect whether word learning efficiency increases as the language proficiency increases.

With these limitations, the current study neither supports nor speaks against the positive effects of L2 proficiency in novel word acquisition. An association between L2 proficiency and word learning efficiency might exist especially in higher proficiency levels, but the current study found little evidence for it. Instead, a few potential caveats in trying to elucidate this question were found. Even though the current study demonstrated that higher L1 proficiency is linked to higher explicit word form and word meaning knowledge, the extent to which implicit word knowledge might be associated with language proficiency still needs clarification. Finally, the current study looked at relatively early stages of novel word acquisition. In order to discover the practical benefits of contextual novel word acquisition, future research should address the question of novel word retention for longer intervals when words have been learnt from context.

4.5 Conclusion

The aim of the current study was to investigate the nature of lexical representations acquired via contextual learning in the auditory domain. The secondary goal was to investigate the effects of L1 and L2 proficiency in novel word acquisition. The research questions set for the study are answered as follows:

1. Can novel word acquisition take place via contextual inference from aurally presented sentences?

Yes. A reliable learning of the novel words was found both in terms of word form recognition accuracy and explicit knowledge of the novel word meanings. This explicit knowledge of the word forms and meanings increased over time, as indexed by higher response accuracy 48 hours after the initial learning of the novel words.

2. Does contextual novel word acquisition in auditory domain result in lexicalization of the novel words and if so, what is the timeline for this process?

Yes, although this wasn't demonstrated with all the used measures of lexicalization. At the level of electrophysiological responses, the novel words could be seen as lexicalized 48 hours after learning: the novel words elicited a semantic priming effect (the N400 effect) when paired with their meanings and crucially also when paired with their semantic associates. The interpretation of the latter effect is that the novel words could activate a wider set of lexical items in their semantic field and this effect cannot be reduced to mere association between the novel word form and its meaning. However, lexicalization effects were not seen at behavioral level in reaction times. These findings are mostly in line with the complementary systems account of word learning: explicit word knowledge almost immediately after learning and later lexicalization effect of meaning of the novel words. In terms of lexicalization of word forms, it is likely that the current procedure failed to detect the effect, rather than an alternative explanation where participants had formed a sufficiently strong link between the meaning and the word form but not integrated the form of the novel words to the mental lexicon. Given the reported early semantic lexicalization effects as indexed by the N400 effect (Mestres-Misse et al., 2007; Perfetti et al., 2005), it is possible that the evidence of semantic lexicalization in the current study reflects an early emergence of lexicalization, before it is detectable on behavioral level.

3. Does acquisition of novel words from context vary as a function of first language or second language proficiency?

Yes. Higher L2 proficiency was found to be associated with higher gains in word form recognition accuracy over time, whereas higher L1 proficiency was associated with higher explicit knowledge of word forms as well as word meanings both immediately and 48 hours after the initial learning of the novel words. This association is interpreted as an auditory equivalent of The Matthew effect observed in reading, although other, potentially complementary interpretations remain possible.

4. Does lexicalization of novel words learnt from context vary as a function of first or second language proficiency?

It seems not. The current data does not support this idea as no language proficiency effects were found in the observed semantic lexicalization (the N400 effect) of novel words. However, it is possible that the proficiency profiles of the participants in the current study were not sufficiently varied to see these effects, especially if the N400 responses in the current study are seen as an early sign of lexicalization of the novel words. More spread in the language proficiency of the participants combined with more stable lexicalization effects (i.e. established by giving more time for memory consolidation or an easier learning task) might show an effect of language proficiency in lexicalization of novel words.

In conclusion, novel word acquisition remains a complex topic with more factors affecting it than what was possible to take into account in the current investigations. The current study adds to a very sparse body of knowledge in the contextual novel word acquisition in the auditory domain. As a part of life long vocabulary development, this mode of novel word acquisition and factors affecting its efficiency requires more research. The role of language proficiency in novel word acquisition has attracted more attention, but robust findings especially in the effects of L2 proficiency are still lacking. This is likely a result of very varied groups of bilinguals and L2 learners used in the previous studies, different measures of L2 proficiency used and different learning and testing procedures utilized. Studies with more unified approach and longitudinal designs would be an important addition to the body of knowledge currently available in this area.

References

Akaike, H. (1985) Prediction and entropy. In A.C. Atkinson and S.E. Fienberg (Ed.), *A Celebration of Statistics* (pp. 1–24). Springer-Verlag: New York.

Alavi, S. M., & Akbarian, I. (2008). Validating a self-assessment Questionnaire on Vocabulary Knowledge. *TELL*, *2*, 125-154.

Astésano, C., Besson, M., & Alter, K. (2004). Brain potentials during semantic and prosodic processing in French. *Cognitive Brain Research*, 18(2), 172–184.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412.

Baddeley, A. D., Gathercole, S. E. & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, 105, 158–173.

Bartolotti, J., & Marian, V. (2012). Language learning and control in monolinguals and bilinguals. *Cognitive Science*, *36*(6), 1129–1147.

Bates, D. M. (2010). lme4: Mixed-effects modeling with R. Retrieved from http://lme4.r-forge.r-project.org/book/

Beck, I., McKeown, M., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. New York: Guilford Press.

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiol*ogy, 60(4), 343–355.

Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. A. (2008). Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes*, 45(2), 122–159.

Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference: A practical information-theoretic approach (2nd Ed.). New York: Springer-Verlag.

Cain, K., Oakhill, J., & Bryant, P. (2004). Individual differences in the inference of word meanings from context: the influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology*, 96(4), 671–681.

Clay, F., Bowers, J. S., Davis, C. J., & Hanley, D. A. (2007). Teaching adults new words: The role of practice and consolidation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5), 970–976.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6), 407- 428.

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering,* 14(3), 391–403.

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: neural and behavioural evidence. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1536), 3773-3800.

De Groot, A. M. B., & Poot, R. (1997). Word translation at three levels of proficiency in a second language: The ubiquitous involvement of conceptual memory. *Language Learning*, 47(2), 215–264.

Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science*, 18(1), 35–39.

Dumay, N., Gaskell, M. G., & Feng, X. (2004). A day in the life of a spoken word. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 339 –344). Mahwah, NJ: Erlbaum.

Durso, F. T., & Shore, W. J. (1991). Partial knowledge of word meanings. *Journal of Experimental Psychology: General*, *120*(2), 190-202.

Edele, A., Seuring, J., Kristen, C., & Stanat, P. (2015). Why bother with testing? The validity of immigrants' self-assessed language proficiency. *Social Science Research*, 52, 99-123. doi:10.1016/j.ssresearch.2014.12.017

Elgort, I., Perfetti, C. A., Rickles, B., & Stafura, J. Z. (2015). Contextual learning of L2 word meanings: second language proficiency modulates behavioural and event-related potential (ERP) indicators of learning. *Language, Cognition and Neuroscience,* 30(5), 506–528.

Ewers, C. A., & Brownson, S. M. (1999). Kindergarteners' vocabulary acquisition as a function of active vs. passive storybook reading, prior vocabulary, and working memory. *Reading Psychology*, 20(1), 11–20.

Fischler, I. (1977). Semantic facilitation without association in a lexical decision task. *Memory & Cognition*, 5(3), 335-339.

Frishkoff, G.A., Perfetti, C.A. & Collins-Thompson, K. (2010). Lexical quality in the brain: ERP evidence for robust word learning from context. *Developmental Neuropsychology*, 35(4), 376–403.

Gaskell, M. G. & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition*, 89(2), 105-132.

Gathercole, S. E., & Baddeley, A. D. (1996). *The Children's Test of Nonword Repetition*. London: Psychological Corporation.

Gathercole, S.E., Frankish, C. R., Pickering, S. J., Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 25, 84–95.

Gupta, P. (2003). Examining the relationship between word learning, nonword repetition, and immediate serial recall in adults. *The Quarterly Journal of Experimental Psychology, 56A,* 1213–1236.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*(5669), 438-441.

Henderson, L., Devine, K., Weighall, A. & Gaskell, G. (2015). When the daffodat flew to the intergalactic zoo: off-line consolidation is critical for word learning from stories. *Developmental Psychology*, 51(3), 406–417.

Holocomb, P. J., & Neville, H. J. (1990). Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Language & Cognitive Processes*, 5(4), 281-312.

Izura, C., Cuetos, F., & Brysbaert, M. (2014). Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicologica*, 35, 49-66.

Jenkins, J. R., Stein, M. L., & Wysocki, K. (1984). Learning vocabulary through reading. *American Educational Research Journal*, 21(4), 767-787.

Kan, P. F., & Sadagopan, N. (2014). Novel word retention in bilingual and monolingual speakers. *Frontiers in Psychology*, 5, 1024. doi:10.3389/fpsyg.2014.01024.

Kan, P. F., Sadagopan, N., Janich, L., & Andrade, M. (2014). Effects of speech practice on fast mapping in monolingual and bilingual speakers. *Journal of Speech, Language, and Hearing Research*, *57(3)*, 929–941.

Kappenman, E. S., & Luck, S. J. (2010). The effects of electrode impedance on data quality and statistical significance in ERP recordings. *Psychophysiology*, 47(5), 888-904.

Kaushanskaya, M. (2012). Cognitive mechanisms of word learning in bilingual and monolingual adults: The role of phonological memory. *Bilingualism: Language and Cognition*, *15*(3), 470–489.

Kaushanskaya, M., & Marian, V. (2008). Age-of-acquisition effects in the development of a bilingual advantage for word learning. *Proceedings of the 32nd Annual Boston University Conference on Language Development*, 213-224. Cascadilla Press; Somerville, MA.

Kaushanskaya, M., & Marian, V. (2009). The bilingual advantage in novel word learning. *Psychonomic Bulletin & Review*, *16*(4), 705–710.

Kaushanskaya, M., & Rechtzigel, K. (2012). Concreteness effects in bilingual and monolingual word learning. *Psychonomic Bulletin & Review*, 19(5), 935–941.

Kaushanskaya, M., Yoo, J., & Van Hecke, S. (2013).Word learning in adults with second language experience: Effects of phonological and referent familiarity. *Journal of Speech, Language, and Hearing Research*, 56(2), 667–678.

Kliegl, R., Masson, M. E. J., & Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition,* 18(5), 655-681.

Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *Modern Language Journal*, 78(3), 285-299.

Kutas, M., & Federmeier, K. D. (2009). N400. *Scholarpedia*, 4, 7790.

Kutas, M., & Federmeier K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology,* 62, 621-647.

Kutas, M., & Hillyard, S. A.(1980) Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427), 203–5.

Kuznetsova A., Brockhoff P. B., & Christensen R. H. B. (2017). LmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), 1-26. doi: 10.18637/jss.v082.i13.

Lindsay, S., & Gaskell, M. G. (2010). A complementary systems account of word learning in L1 and L2. *Language Learning*, 60, 45-63.

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review*, 7(4), 618–630.

Luke, S. G. (2016). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, 49(4), 1494-1502.

Majerus, S., Poncelet, M., Van der Linden, M., & Weekes, B. S. (2008). Lexical learning in bilingual adults: The relative importance of short-term memory for serial order and phonological knowledge. *Cognition*, 107(2), 395–419.

Majerus, S., Van der Linden, M., Mulder, L., Meulemans, T., & Peters, F. (2004). Verbal short-term memory reflects the sublexical organization of the phonological language network: Evidence from an incidental phonotactic learning paradigm. *Journal of Memory and Language*, 51(2), 297-306.

Marian, V., & Spivey, M. (2003). Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition, 6(2)*, 97–115.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, 25(1-2), 71–102.

Mattys, S.L., & Clark, J.H. (2002). Lexical activity in speech processing: Evidence from pause detection. *Journal of Memory and Language*, 47(3), 343–359.

Mestres-Misse A., Rodriguez-Fornells A., & Münte T. F. (2007). Watching the brain during meaning acquisition. *Cerebral Cortex*, 17(8), 1858–1866.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2), 227-234.

Min, B. K., Park, J. Y., Kim, E. J., Kim, J. I., Kim, J. J., & Park, H. J. (2008). Prestimulus EEG alpha activity reflects temporal expectancy. *Neuroscience Letters.* 438(3), 270–274.

Nagy, W.E., Herman, P.A. & Anderson, R.C. (1985). Learning words from context. *Reading Research Quarterly*, 20, 233-253.

Nair, V. K. K., Biedermann, B., & Nickels, L. (2016). Consequences of late bilingualism for novel word learning: Evidence from Tamil–English bilingual speakers. *International Journal of Bilingualism*, 20(4), 473-487.

Papagno, C., & Vallar, G. (1995). Short-term memory and vocabulary learning in polyglots. *Quarterly Journal of Experimental Psychology*, 48A, 98-107.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1-2), 8–13.

Perfetti C. A., Wlotko E. W., & Hart L. A. (2005). Word learning and individual differences in word learning reflected in event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 31(6),1281–1292.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18*, 22–37.

Pulido, D. (2003). Modeling the role of second language proficiency and topic familiarity in second language incidental vocabulary acquisition through reading. *Language Learning*, 53(2), 233-284.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raven, J. C. (1958) *Standard Progressive Matrices: Sets A, B, C, D and E.* London: Lewis, H. K. & Co. Ltd.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(1), 26-43.

Sénéchal, M., Thomas, E. & Monker, J. A. (1995). Individual differences in 4-year-old children's acquisition of vocabulary during storybook reading. Journal of Educational Psychology, 87(2), 218 –229.

Shefelbine, J. L. (1990). Student factors related to variability in learning word meanings from context. *Journal of Reading Behavior*, 22, 71–97.

Staehr, L. S. (2008). Vocabulary size and the skills of reading, listening and writing. *Language Learning Journal*, 36(2), 139-15.

Stanovich, K. E. & Cunningham, A. E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Memory and Cognition*, 20, 51-68.

Steinhauer, K., Alter, K., & Friederici, A. D. (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience*, 2(2), 191–196.

Sternberg, R. J. (1987). Most vocabulary is learned from context. In M.G. McKeown & M.E. Curtis (Eds.), *The Nature of Vocabulary Acquisition* (p. 89-105). Hillsdale, N.J.: Lawrence Erlbaum.

Swanborn, M. S. L., & De Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. Review of Educational Research, 69(3), 261-286.

Tabachnick, B. G., & Fidell, L. S. (2001). Using multivariate statistics (4th ed.). Needham Heights, MA: Allyn & Bacon.

Tamminen, J. & Gaskell, M. G. (2013).Novel word integration in the mental lexicon: Evidence from unmasked and masked semantic priming. *Quarterly Journal of Experimental Psychology*, 66(5), 1001-1025.

Tannenbaum, K. R., Torgesen, J. K., & Wagner, R. K. (2006). Relationships between word knowledge and reading comprehension in third-grade children. *Scientific Studies of Reading*, 10(4), 381–398.

van der Ven, F., Takashima, A., Segers, E., & Verhoeven, L. (2015). Learning word meanings: Overnight integration and study modality effects. *PLoS One*, 10(5), e0124926.

Van Hell, J. G., & Mahn, A. C. (1997). Keyword mnemonics versus rote rehearsal: Learning concrete and abstract foreign words by experienced and inexperienced learners. *Language Learning*, 47(3), 507-546.

Wechsler, D. (1997). *WAIS-III Administration and scoring manual*. San Antonio, TX: The Psychological Association.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers, 20,* 6–11. Web: http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm

Appendix A – Properties of the stimuli used in the experiment.

Table A1. List of Base word, Novel word (non-word) and Foil triplets in experimental lists A and B. The pronunciation of Novel words and Foils matched those of Base words up until the final vowel and consonant(s), where the Novel words and Foils deviated from the Base words. Reduced vowels in the deviating section of items are transcribed as / ə /.

List A			List B		
Base word	Novel word	Foil	Base word	Novel word	Foil
artichoke	artichəd	artichən	alcohol	alcohin	alcohid
avalanche	avalogue	avalot	anecdote	anecdəl	anecdən
badminton	badmintel	badmintet	antelope	anteluce	anteluke
blossom	blossail	blossain	assassin	assassool	assassood
canopy	canopule	canopute	bayonet	bayoniss	bayonil
capsule	capsoth	capsod	boulevard	boulevett	bouleven
caravan	caravoth	caravol	bracelet	bracelang	bracelar
cardigan	cardigite	cardigile	canteen	cantove	cantode
casket	caskal	caskan	cartridge	cartroce	cartrole
casserole	casserin	casserith	cellophane	cellophoke	cellophoce
cathedral	cathedruke	cathedruce	cinnamon	cinnamil	cinnamig
cucumber	cucumbeat	cucambeak	citadel	citadin	citadist
decibel	decibon	decibob	clarinet	clarinern	clarinerl
diaphragm	diaphrume	diaphrude	culprit	culpran	culprass
emperor	emperan	emperaph	detergent	detergile	detergice
fugitive	fugitein	fugiteid	dialogue	dialaiff	dialaist
galaxy	galaxum	galaxuff	dungeon	dungell	dungeck
helium	heliac	heliat	gelatine	gelatord	gelatorl
incentive	incentar	incentark	gorilla	gorillin	gorillit
lantern	lantobe	lantoke	hamster	hamstoch	hamstol
mackerel	macerine	mackerife	hemisphere	hemisphed	hemisphen
monsoon	monsteen	monsteece	hurricane	hurricarb	hurricarth
parachute	parashəff	parashən	kangaroo	kangariff	kangarin
pavilion	paviliate	paviliage	molecule	molekyən	molekyək
porcelain	porcelote	porcelobe	napkin	napkəm	napkəss
pyramid	pyramon	pyramotch	ornament	ornameast	ornameab
souvenir	souvenart	souvenark	parsnip	parsnəg	parsnəs
squirrel	squirrome	squirrope	pelican	pelikive	pelikibe
surplus	surplode	surplone	skeleton	skeletobe	skeletope
tulip	tulode	tulome	slogan	slowgiss	slowgith
vendetta	vendetrick	vendetrip	stamina	stamingent	stamingelk
vestibule	vestibate	vestibain	vinegar	vinegate	vinegale

Table A2. Stimuli used in the Semantic Relatedness task in Real (R), Meaning (M) and Semantic Associate (SA) conditions. Novel word form (Prime in M and SA conditions) and related and unrelated targets in each condition are shown with Δ Phon. (Difference in phonemic length between related and unrelated targets) and Δ Freq. (Difference in frequency of occurrence between related and unrelated targets). Differences are expressed in absolute values.

Nove word form		Targets in M condition				Targets in R and SA conditions			
List A	List B	Related	Unrelated	Δ Phon.	Δ Freq.	Related	Unrelated	Δ Phon.	Δ Freq.
badmintel	molecule	airport	hammer	0	49	terminal	freezer	1	10
cathedruke	boulevett	basket	camera	1	12	weave	nail	0	23
vestibate	detergile	bicycle	winter	1	60	pedal	suite	0	10
incentar	clarinern	bottle	clock	0	77	glass	pouch	1	140
paviliate	alcohin	camera	basket	1	12	photograph	ambulance	0	51
cucumbeat	skeletobe	casino	helmet	0	7	roulette	terminal	1	12
galaxum	bracelang	ceiling	needle	1	14	floor	thumb	0	149
fugitein	ornameast	cigarette	dentist	0	62	ashtray	tractor	1	2
caravoth	cinnamil	clock	bottle	0	77	wristwatch	nostril	0	10
pyramon	stamingent	dentist	cigarette	0	62	tooth	floor	0	88
macerine	anteluce	driver	finger	0	67	taxi	salt	1	3
diaphrume	hemisphed	engine	pocket	0	14	petrol	headline	1	2
casserin	hurricarb	farm	nose	0	4	tractor	humor	1	13
capsoth	culpran	finger	driver	0	67	thumb	sun	0	125
parasheff	anecdel	fridge	lion	0	21	freezer	petrol	0	12
avalogue	kangariff	hammer	airport	0	49	nail	armour	0	15
vendetrick	gelatord	helmet	casino	0	7	armour	weave	0	8
cardigite	pelikive	hospital	newspaper	1	7	ambulance	photograph	0	51
articheD	cellophoke	hotel	scissors	0	139	suite	thread	0	4
decibon	dialaiff	joke	sugar	1	7	humor	roulette	0	22
emperan	vinegate	lion	fridge	0	21	roar	sock	1	7
surplode	slowgiss	map	shoe	1	39	compass	ashtray	1	4
blossail	cantove	moon	phone	0	3	sun	tooth	0	64
squirrome	napkem	needle	ceiling	1	14	thread	taxi	1	18
heliac	citadin	newspaper	hospital	1	7	headline	compass	0	9
caskal	dungell	nose	farm	0	4	nostril	wristwatch	0	10
souvenart	gorillin	phone	moon	0	3	dial	summer	0	120
lantobe	hamstoch	pocket	engine	0	14	pouch	glass	1	140
tulode	parsneg	scissors	hotel	0	139	barber	dial	0	3
porcelote	assassool	shoe	map	1	39	sock	roar	1	7
canopule	cartroce	sugar	joke	1	7	salt	pedal	0	35
monsteen	bayoniss	winter	bicycle	1	60	summer	barber	0	117
Average				**0.4**	**36.4**			**0.4**	**40.1**
Minimum				**0**	**3**			**0**	**2**
Maximum				**1**	**139**			**1**	**149**

Table A3. Key properties of the sentence contexts in which the novel words were learnt from. Mean number of words in a context and mean cloze probability show the mean for each set of 6 sentences for each novel word with minimum and maximum in parentheses. Replaceability in learning phase shows the percentage of sentences in a set of 6 that could make sense using the semantic associate of the meaning of the novel word. Replaceability in restudy phase shows a yes/no answer to whether the single sentence presented for each novel word in the restudy phase could have been answered correctly while interpreting the sentence with the semantic associate rather than the meaning of the novel word.

Novel word meaning in the context	Mean no. of words (Min - Max)	Mean Cloze Probability (Min - Max)	Replaceability in learning phase (%)	Replaceability in restudy phase
Airport	15.5 (14 - 19)	85.3 (76 - 100)	100	Yes
Basket	13.8 (12 - 15)	93.4 (92 - 100)	0	No
Bicycle	12.8 (10 - 14)	86.7 (76 - 100)	0	No
Bottle	16.7 (14 - 19)	92 (84 - 100)	33.3	No
Camera	13.5 (11 - 18)	89.3 (76 - 100)	16.7	No
Casino	13.8 (12 - 15)	90.7 (84 - 100)	33.3	No
Ceiling	14.5 (13 - 18)	89.2 (84 - 100)	0	No
Cigarette	13.2 (11 - 17)	84.3 (69 - 100)	33.3	No
Clock	16.2 (12 - 19)	97.3 (92 - 100)	33.3	No
Dentist	14 (13 - 17)	92 (76 - 100)	0	No
Driver	15.5 (14 - 19)	96 (92 - 100)	0	No
Engine	14.2 (12 - 17)	86.8 (69 - 100)	0	No
Farm	15.3 (14 - 17)	88 (84 - 100)	0	No
Finger	15.5 (13 - 17)	94.7 (84 - 100)	16.7	Yes
Fridge	14.3 (13 - 16)	88.2 (69 - 100)	66.7	No
Hammer	15.3 (11 - 20)	85.7 (69 - 100)	0	No
Helmet	14.8 (13 - 16)	86.8 (69 - 100)	33.3	No
Hospital	14.6 (12 - 17)	94.7 (84 - 100)	0	No
Hotel	14.8 (11 - 17)	96 (92 - 100)	50	Yes
Joke	15.2 (13 - 17)	90.7 (84 - 100)	16.7	No
Lion	15.2 (11 - 18)	88 (76 - 100)	0	No
Map	16.3 (14 - 19)	93.3 (84 - 100)	50	Yes
Moon	14.5 (11 - 17)	96 (84 - 100)	0	No
Needle	16.5 (14 - 22)	93.3 (76 - 100)	50	No
Newspaper	14 (12 - 19)	93.3 (84 - 100)	0	No
Nose	13 (11 - 15)	98.7 (92 - 100)	33.3	Yes
Phone	15.2 (13 - 17)	93.3 (84 - 100)	0	No
Pocket	15.5 (13 - 18)	97.3 (84 - 100)	50	Yes
Scissors	15.8 (13 - 19)	92 (84 - 100)	0	No
Shoe	14.7 (11 - 18)	94.7 (84 - 100)	16.7	Yes
Sugar	15.5 (12 - 18)	89.5 (69 - 100)	16.7	No
Winter	16.7 (13 - 20)	90.8 (69 - 100)	16.7	No
Total of all contexts	**14.9 (10 - 22)**	**91.5 (69 - 100)**	**20.8**	**7 Yes**

Appendix B – Response accuracies in the Semantic Relatedness task by condition and trial type.

Appendix B1. Correct responses (%) for Real (R), Meaning (M) and Semantic Associate (SA) conditions in the Semantic Relatedness task for Related and Unrelated trials. Mean accuracies are shown with standard deviation in brackets. < 50 % indicates the number of participants who had less than 50% correct responses in any given condition-trial type combination.

Session 1	R condition		M condition		SA condition	
	Related	Unrelated	Related	Unrelated	Related	Unrelated
Mean	95.6 (3.5)	96.8 (2.9)	72.2 (18.1)	82.4 (12.3)	61 (20)	81.8 (12.7)
Min	87.5	87.5	25	50	21.9	50
Max	100	100	96.9	100	93.8	96.9
< 50 %	0	0	2	0	7	0

Session 2	R condition		M condition		SA condition	
	Related	Unrelated	Related	Unrelated	Related	Unrelated
Mean	95.4 (3.7)	98.4 (1.9)	75.3 (17.5)	86.2 (12.7)	66.8 (18.7)	86.3 (13.4)
Min	87.5	93.8	43.8	50.0	21.9	46.9
Max	100	100	100	100	93.8	100
< 50 %	0	0	3	0	5	1

Appendix C – Analysis of the ERP data for Session 1 and Session 2 separately in the 300-600 ms and 600-900 ms time windows.

The best fit for the data from both sessions (Model-36) in the 300-600 ms time window was fitted for the data from Session 1 and Session 2 separately to investigate the emergence of the N400 effect in more detail. In these analyses, Model-36 will be referred to as Model-bySession for clarity, as the same model is used for data in the 600-900 ms time window, which is described below. The R formula for the Model-bySession is as follows:

Model-bySession: ERP.300-600 ~ Condition*TrialType + TrialType + Condition + (1|Subject) + (1|Item)

The separate analyses for Session 1 and Session 2 yielded model estimates that were nearly a perfect match to the observed data (see table C1). The best models for Session 1 and Session 2 data were chosen based on AICc values as described in Model selection Procedure. The best model for Session 1 data was Model-bySession. Whereas Model-bySession was the second best fit for Session 2 data only (with $\Delta AICc < 2$)[7], the best fit for Session 2 data was a model with fixed effects of Condition and TrialType as predictors (referred to as Model-Session2). However, the estimates for the TrialType difference from Model-bySession were still superior to the estimates from Model-Session2 (see Table C1). Additionally, the pattern of results suggested by both models is largely comparable. As such, the Model-bySession is considered the best fit for both Session 1 and Session 2 data. See Table C2 for a full list of candidate models and AICc values in the model selection procedure for data from Session 1 and Session 2.

[7] When the difference in AICc values ($\Delta AICc$) between candidate models is less than 2, such a small AICc difference indicates that the candidate model with slightly higher AICc value still has substantial empirical support, given the data (Burnham & Anderson, 2002, p.70).

Table C1. Observed values compared with Model estimates for difference in ERP responses to related vs unrelated trials (TrialType difference) in Real word, Meaning and Semantic Associate conditions. Model estimates of Model-bySession (Condition*TrialType + TrialType + Condition) fitted for data from Session1 and Session 2 separately in 300-600 ms time window and model estimates of Model-Session2 (TrialType + Condition) fitted for data from Session 2 only in 300-600 ms time window.

Condition	Session	Observed value	Model-bySession estimate	Model-Session2 estimate
Real	1	-1.09	-1.09	
Meaning	1	-0.46	-0.47	
Semantic Associate	1	-0.37	-0.38	
Real	2	-1.12	-1.12	-0.93
Meaning	2	-1.04	-1.04	-0.93
Semantic Associate	2	-0.63	-0.63	-0.93

*Table C2.*Candidate models for ERP analyses for Session 1 and Session 2 in 300-600 ms time window. Models are in ascending order based on AICc values with the best fitting model in bold. ΔAICc indicates the difference in AICc values between the best fitting model and another candidate model.

Session 1

Fixed effects of the candidate model, 300-600 ms	df	AICc	ΔAICc
Condition * TrialType + TrialType + Condition	**9**	**27130.94**	**0**
Condition + TrialType	7	27133.73	2.79
TrialType	5	27144.56	13.62
Condition * TrialType * L1 + TrialType + Condition + L1	15	27154.13	23.19
Condition	6	27165.17	34.23
no fixed effects	4	27175.82	44.88
Condition * TrialType * L2 + TrialType + Condition + L2	15	27184.05	53.11

Session 2

Fixed effects of the candidate model, 300-600 ms	df	AICc	ΔAICc
Condition + TrialType	**7**	**27393.02**	**0**
Condition * TrialType + TrialType + Condition	9	27394.74	1.72
TrialType	5	27398.95	5.93
Condition * TrialType * L1 + TrialType + Condition + L1	15	27428.27	35.25
Condition * TrialType * L2 + TrialType + Condition + L2	15	27453.2	60.18
Condition	6	27472.54	79.52
no fixed effects	4	27478.32	85.3

The same model selection procedure was followed for data from Session 1 and Session 2 in the 600-900 ms time window. Model-bySession was the best fit for data from both sessions (see Table C3).

*Table C3.*Candidate models for ERP analyses for Session 1 and Session 2 in 600-900 ms time window. Models are in ascending order based on AICc values with the best fitting model in bold. ΔAICc indicates the difference in AICc values between the best fitting model and another candidate model.

Session 1

Fixed effects of the candidate model, 600-900 ms	df	AICc	ΔAICc
Condition * TrialType + TrialType + Condition	**9**	**28260.59**	**0**
Condition + TrialType	7	28266.24	5.65
Condition	6	28288.84	28.25
Condition * TrialType * L1 + TrialType + Condition + L1	15	28291.84	31.25
Condition * TrialType * L2 + TrialType + Condition + L2	15	28315.57	54.98
TrialType	5	28422.04	161.45
no fixed effects	4	28443.71	183.12

Session 2

Fixed effects of the candidate model, 600-900 ms	df	AICc	ΔAICc
Condition * TrialType + TrialType + Condition	**9**	**28440.91**	**0**
Condition + TrialType	7	28443.12	2.21
Condition * TrialType * L1 + TrialType + Condition + L1	15	28477.49	36.58
TrialType	5	28485.87	44.96
Condition	6	28486.81	45.9
Condition * TrialType * L2 + TrialType + Condition + L2	15	28495.88	54.97
no fixed effects	4	28529.14	88.23