Interactive L2 vocabulary acquisition in a lab-based immersion setting

Johanna F. de Vos, Herbert Schriefers, Louis ten Bosch & Kristin Lemhöfer

View supplementary material 🗗

Published online: 09 Apr 2019.

Submit your article to this journal 🗗

Article views: 96

View Crossmark data 🗗

Routledge
Taylor & Francis Group

REGULAR ARTICLE

Interactive L2 vocabulary acquisition in a lab-based immersion setting

Johanna F. de Vos[a,b], Herbert Schriefers[a], Louis ten Bosch[a,c] and Kristin Lemhöfer[a]

[a]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands; [b]International Max Planck Research School for Language Sciences, Nijmegen, The Netherlands; [c]Centre for Language Studies, Radboud University, Nijmegen, The Netherlands

ABSTRACT
We investigated to what extent L2 word learning in spoken interaction takes place when learners are unaware of taking part in a language learning study. Using a novel paradigm for approximating naturalistic (but not necessarily non-intentional) L2 learning in the lab, German learners of Dutch were led to believe that the study concerned judging the price of objects. Dutch target words (object names) were selected individually such that these words were unknown to the respective participant. Then, in a dialogue-like task with the experimenter, the participants were first exposed to and then tested on the target words. In comparison to a no-input control group, we observed a clear learning effect especially from the first two exposures, and better learning for cognates than for non-cognates, but no modulating effect of the exposure-production lag. Moreover, some of the acquired knowledge persisted over a six-month period.

Introduction

In 2015, almost a quarter billion people were living abroad as immigrants, and their numbers are rising (United Nations, 2015). For the majority of these people, moving to a new country means moving to a second language (L2) environment. While some people fully rely on immersion in the L2 environment for developing their language skills and building a new vocabulary, others start out by taking language classes. But in the end, even those who were tutored for a while will likely end up growing most of their L2 vocabulary knowledge through daily-life interactions with native speakers of the target language.

In this study, we investigated what vocabulary acquisition in immersed L2 interaction looks like, starting from the moment when learners hear a word that they did not know before. How quickly can they acquire such new words, and does this knowledge persist over time? For the first time, these questions were addressed in an experimental setting, whose aim (i.e. L2 word learning) was fully hidden from the participants. This was done in the hope that any resulting learning would be the best approximation of naturalistic L2 learning that can be obtained in a laboratory.

Immersion and incidental learning

There are two large research strands that touch upon different aspects of the above questions, but neither fully answers it. The first strand, L2 immersion research, investigates the language skills and language development of learners who live, work and/or study in an L2 environment. Unsurprisingly, learners who have been immersed longer, and/or to a higher degree, generally score better on measures of L2 lexical proficiency, for example on lexical categorisation (e.g. Malt & Sloman, 2003; Zinszer, Malt, Ameel, & Li, 2014) and receptive vocabulary (e.g. Dahl & Vulchanova, 2014).

In the current study, we strove to simulate an L2 immersion setting in the lab and apply various experimental manipulations within that context. In other words, we aimed to observe learning as it happens during immersion, rather than to compare learning between learners who differ in the extent or duration of their L2 immersion, as was done in the studies described above. Such studies would typically be non-experimental, because learners usually are not assigned to different degrees of immersion (one exception is Dahl & Vulchanova, 2014). Other studies have also focused on learning within an immersion setting (e.g. Lapkin, Swain, & Smith, 2002; Swain & Lapkin, 1995), but these studies were conducted in L2 classrooms. In those cases, it can be expected that more of the learners' attention was devoted to L2 word learning than would be the case in daily life.

The second research strand is that of incidental word learning. This strand also investigates vocabulary

---

acquisition in interactions that are not explicitly aimed at word learning. A review of definitions, potential mechanisms and operationalisations of incidental learning is given in De Vos, Schriefers, Nivard, and Lemhöfer (2018). In summary, incidental learning is often defined in one of three ways. The first revolves around the learners' intentions: Incidental learning would be "learning without intention, while doing something else" (Ortega, 2009, p. 94). This definition is intuitively appealing, but intentions are hard to measure and may also change over time. Easier to operationalise is the second definition: whether or not an upcoming post-test is announced to the learners (Hulstijn, 2003). The third definition revolves around the activity that the learners engage in: For learning to be incidental, it should come about as a "by-product" (Hulstijn, 2003, p. 362) of a task that primarily revolves around meaning.

There is a long and rich research tradition in incidental learning, which has investigated many variables that potentially influence the degree of learning, and that may also be relevant to the current topic. Examples of such variables are the number of exposures to a new word (e.g. Godfroid et al., 2018; Gullberg, Roberts, & Dimroth, 2012; Van Zeeland & Schmitt, 2013), the text genre (e.g. Shokouhi & Maniati, 2009), the context that a word appears in (e.g. Bordag, Kirschenbaum, Tschirner, & Opitz, 2015; Vidal, 2011), and individual differences (e.g. Grey, Williams, & Rebuschat, 2015; Robinson, 2002). For review articles on incidental L2 word learning, see De Vos, Schriefers, Nivard, et al. (2018), Ellis (1999), Huckin and Coady (1999), Hulstijn (2003), Restrepo Ramos (2015) and Schmitt (2008).

Especially when incidental learning is operationalised according to the second and third definitions, it appears to be related to the kind of learning we are interested in (i.e. naturalistic learning). However, the existing research is typically conducted in contexts that are quite explicitly geared towards L2 learning, which sets these learning contexts apart from the ones that learners usually encounter in their daily lives. The majority of incidental L2 learning studies are conducted in non-immersive L2 classrooms in the home country of the participants. Even if a school uses an immersion programme, the learners will obviously know that the activities in the L2 classroom are aimed at improving their language skills.

Studies on incidental learning are also sometimes conducted in labs, which removes the focus on L2 learning that is inevitable in the L2 classroom. For example, McGraw, Yoshimoto, and Seneff (2009) recruited students from American universities with at least one semester of Mandarin experience to take part in a lab-based study. The participants played interactive card games, in which they incidentally encountered Mandarin

words. Gullberg et al. (2012) recruited Dutch students with no prior experience with Mandarin, and let them watch a Mandarin weather report video. These participants were not informed of the researchers' interest in vocabulary, nor did they know that they would later be tested on Mandarin vocabulary. Still, in both studies the participants must have been aware of participating in a language-related experiment – why would they otherwise be exposed to Mandarin and recruited based on their Mandarin experience?

The conclusion from the incidental learning literature so far is that it has not provided insight in naturalistic L2 word learning in an immersion setting, because the research has mainly been situated in contexts which obviously revolved around L2 learning. In many of the existing studies, the participants could draw these conclusions from being tested in an L2 classroom, in a novel or foreign language different from the language in their environment, or from being recruited based on their language background. The administration of vocabulary pre-tests could also add to the suspicion that a study may concern language learning, and that a post-test could follow. Although the above review has focused on learning from spoken rather than written input, the same arguments generally apply to studies on incidental L2 word learning from reading. As it can be expected that participants approach experimental activities from a different angle when they suspect they should be learning words, there is a need for research that better approximates real-life interactive L2 learning in an immersion setting by hiding the study's language learning aspect.

One such study was conducted by De Vos, Schriefers, and Lemhöfer (2018), who investigated the effects of noticing vocabulary "holes" on subsequent L2 incidental word learning. Having a vocabulary hole (Doughty & Williams, 1998) means having no knowledge of a particular word; noticing a vocabulary hole means to become aware of this lack of knowledge. This contrasts with the more commonly used term *noticing the gap* (Schmidt & Frota, 1986), which describes the situation in which learners become aware of the discrepancy between how they are using a certain word or structure, and the way it is used by a more proficient or native speaker of the target language.

The participants in De Vos, Schriefers, and Lemhöfer (2018) were German native speakers living in the Netherlands who did not know they had been recruited based on their language background. They took part in a task which they were told revolved around comparing objects by price. In reality, however, it was investigated whether the participants would learn the objects' names. It was found that the participants who had

previously noticed vocabulary holes on average were able to recall more words than those participants who had not. Most important with regard to the present study is that De Vos, Schriefers, and Lemhöfer (2018) showed that their price judgment task worked extremely well for disguising the learning aim of the study.

The present study

The present study used a similar experimental set-up as De Vos, Schriefers, and Lemhöfer (2018), but was new in the fact that the participants this time not only listened to native-speaker input, but also produced the L2 target words in alternation with the experimenter. This comes closer to taking part in real-life conversational settings. Of course, we acknowledge that a lab-based study can never be fully representative of real-life naturalistic language learning. On the other hand, the experimental control that comes with lab-based studies allowed us to take into account the participants' pre-existing productive knowledge of the target words, and to select target items accordingly on an individual basis for each participant. This approach, used here for the first time, enabled us to work with natural language items (as opposed to pseudowords), making the study more realistic, while still ensuring that all participants actively learned an equal number of previously (productively) unknown words. Furthermore, we could exactly control the input the participants were exposed to during the experiment, including when and how often the target words were presented.

The study was advertised as a psychological experiment about making price judgments. Of actual interest to us, however, was to what extent the German participants would learn to produce the Dutch names of the objects that they compared by price. As our participants already knew Dutch, it was possible that they also had pre-existing knowledge of the target objects' names. Therefore, we conducted a pre-test, but called it a "sorting task" and disguised it as part of the price judgment task. For each participant, the experimental software made a separate selection of target and filler items based on the outcomes of the pre-test. This had the advantage that all participants were exposed to an equal number of Dutch words productively unknown to them (thus, experiencing the same memory load), albeit not necessarily the same words across participants. While the use of artificial language items would have been less complicated, we think that encountering a set of pseudowords that could in no way be linked to one's existing L2 vocabulary would quickly induce participants' suspicion with regard to the study's real purpose.

After the items had been selected, the participant engaged in an interactive task (the "price comparison task") with the experimenter, who was a Dutch native speaker. The participant and the experimenter took turns producing utterances comparing two objects by price. Only for participants in the experimental group did the price judgments made by the experimenter contain the target objects' names. This provided these participants with the opportunity to learn the target words. Whether or not the participants could name these objects in later trials was the dependent variable and the measure of word learning. Twenty minutes and six months after the learning phase, the retention of the target objects' names was tested again with a picture-naming task.

The primary aim of this study was to investigate how many L2 words can be learned under these circumstances, and how much of the newly-acquired knowledge is retained over the course of 20 minutes and six months. In addition, the structured conversational setting also provided the opportunity to investigate the predictors of cognate status, exposure frequency and the lag between exposure and production, which are known to affect memory performance under explicit learning conditions (more details are given below).

How much learning?

Because the current study was the first to investigate interactive L2 word learning in an immersion setting while the participants were unaware of taking part in a language learning study, of primary interest to us were their learning rates. In De Vos, Schriefers, and Lemhöfer (2018), the learners were also unaware, but did not alternate with the experimenter in producing the target words during the learning phase (in other words, the learning in that study was not interactive). Given this difference, we were interested to know how large the learning effect would be in the current setting. In order to correctly estimate the size of the learning effect, we also included a control group that was not exposed to the target words at all, but was still tested on them. This allowed us to separate learning effects from potential testing effects, guessing effects, and spontaneous fluctuations in the participants' behaviour.

Exposure frequency

Another difference to De Vos, Schriefers, and Lemhöfer (2018) was that their participants were only tested after having been exposed to the target words four times. At that point, they scored 28% correct. However, it was unknown how the participants' word knowledge grew depending on the number of exposures they received.

Therefore, in the present study we tested the participants both after two and four exposures to the target words.

It is known that having more exposures to an L2 word generally (although not always) results in better acquisition (e.g. Bisson, Van Heuven, Conklin, & Tunney, 2014; Rott, 1999; Van Zeeland & Schmitt, 2013; Vidal, 2011), but the relationship between exposure frequency and word learning can take different shapes. One possibility is that little learning occurs at first (here, after two exposures), but that substantial learning would be visible after more exposures (here, four). If so, two exposures would apparently not be enough for creating new entries in the L2 mental lexicon, while this threshold could be crossed with four exposures. On the other hand, it is conceivable that two exposures already suffice for learning a new word, and that the third and fourth exposure would not add much. Both types of outcomes are seen in the literature.

For example, Vidal (2011) studied the role of exposure frequency in L2 word learning from reading and listening. The effect of exposure frequency differed per mode: In reading, the greatest gains were found between two and three exposures, while in the case of listening, exposures one to five had very little impact, but there was a steep increase in the scores after six exposures. Bisson et al. (2014) compared two, four, six and eight exposures and found that the first two exposures relatively had a lot of impact on learning rates, while the impact of subsequent exposures decreased and, descriptively, no longer seemed to change between six and eight exposures. Thus, among other things, the relationship between exposure frequency and L2 word learning seems to be dependent on the type of input and other details of the experimental design.

In the present study, we wished to quantify this relationship in the lab-based setting we had created for studying naturalistic, interactive L2 word learning. Exposure frequency was manipulated and tested within words. This seems reflective of real-life conversations, where learners often already try to use new words even if they have not yet mastered them perfectly, and then will subsequently hear these words again. Productive knowledge of the target words was measured after zero exposures (in the pre-test), and after two and four exposures (in the price comparison task). With the term *exposure*, we refer to those moments in which a participant was exposed to a target word in the speech of the experimenter. If a participant correctly produced a target word in one of the measurements in the experiment, one could technically also call that an exposure, but this was not the same for all the participants. In addition, no feedback was given on the correctness of

the participants' target word productions during the price comparison task. For these reasons, we will use the term *exposure* only in reference to the experimenter's use of the target words. We hypothesised that the participants would achieve higher scores after more exposures. We regarded the question of the relative impact of two versus four exposures as an exploratory rather than a hypothesis-based question.

Cognate status

Cognates are L1–L2 translation word pairs that share a common origin, which can still be seen from similarity in form and meaning. Word learning studies conducted under explicit learning conditions have shown that cognates are easier to learn than non-cognates (e.g. Lotto & De Groot, 1998) and are also less susceptible to forgetting (e.g. De Groot & Keijzer, 2000). The facilitative effect of cognate status can both be explained at the stage of word form learning, where there is relatively less new information to be learned, and at the stage of retrieval, where a translation is directly activated due to the phonological similarity between the L1 and L2 word forms (De Groot, 2011, p. 119).

In the studies referenced above, the participants learned cognate and non-cognate words under explicit learning conditions, namely through paired-associate training. In the present study, we tested whether the cognate advantage is also found when learners' attention is not explicitly drawn to word learning. We expected that, in these circumstances, cognates will still benefit from their similarity to existing L1 word form representations.

Exposure-production lag

The retention interval is the time that passes between the final study episode of an item, and the test of this item (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006, p. 354). Typical word learning studies consist of a learning phase and one or multiple post-tests, with the retention interval varying from a few minutes after the learning phase to days, weeks or months (e.g. Brown, Waring, & Donkaewbua, 2008; Van Zeeland & Schmitt, 2013). We are not aware of any studies in which L2 word learning was tested with various retention intervals during the learning phase itself, or in other words, studies in which training and test trials alternate. This is relevant, because in real-life conversations learners often put newly acquired words directly into use rather than wait until the conversation is already over. Therefore, in the current study we tested word learning with short retention intervals, which we will call *lags*, similar to those in real-life conversation (i.e. a few utterances after exposure).

Outside the domain of L2 word learning, there are several studies on L1 paired-associate learning in which test and training trials do alternate. These studies have shown that the second half of a word pair is generally recalled more accurately after a shorter lag (e.g. Balota, Duchek, & Paullin, 1989; Peterson, Wampler, Kirkpatrick, & Saltzman, 1963). However, L1 paired-associate learning with written stimuli is different from interactive L2 word learning when learners are unaware of the study's word learning aspect. Thus, the question arises whether L2 words that are learned during conversation similarly benefit from having a shorter lag (here, three trials) rather than a longer one (here, seven trials).

Long-term retention

In addition, we were also interested in the participants' long-term retention of their newly acquired word knowledge after two different retention intervals: twenty minutes and six months. After all, learners usually want to not only expand their vocabulary for use in the moment, but also for future use. This especially applies to learners who are using the L2 in their daily life, like our participants (in contrast to learners whose main motivation may be getting good grades on a school exam). We chose the 20-minute retention interval partly for practical reasons (so that this first post-test could be administered in the same session), and partly because 20 minutes is a commonly used retention interval in long-term memory studies (e.g. Anderson, Bjork, & Bjork, 1994; Loftus, Miller, & Burns, 1978; MacLeod & Macrae, 2001; Williams & Zacks, 2001). We chose the six-month retention interval to gain insight in forgetting over a very long period of time; this retention interval is longer than is typically found in studies on long-term retention (a few days, weeks or months are the more commonly used retention intervals).

Research questions

The issues raised above can be summarised in the following research questions:

(1) What are the L2 word learning rates from spoken interaction, for immersed learners who are unaware of taking part in a language learning study?
(2) Do vocabulary gains vary as a function of:
   (a) Cognate status? (cognate versus non-cognate)
   (b) Exposure frequency? (two versus four exposures)
   (c) Lag? (three versus seven trials)
(3) How much vocabulary do learners still remember after retention intervals of 20 minutes and six months after the experiment?

Methods

Participants

Sixty-one native speakers of German in Nijmegen, the Netherlands, were recruited for the experiment. They were rewarded with money or course credits. All participants were enrolled in, or had recently graduated from, a Dutch university. In recruitment, care was taken to ensure that participants remained unaware of the study being about L2 learning. The study was advertised as a psychological experiment about making price judgments. Eligibility requirements only mentioned that participants needed to be able to speak Dutch, but did not mention any restrictions with regard to native language. The online participant recruitment system made it possible for us to selectively advertise the study to German native speakers only.

Fifteen participants would later be excluded from the analysis because they had too much pre-existing knowledge of the target words (see Procedure). One additional participant was excluded because she had correctly guessed that the experiment was about L2 word learning. The final sample thus consisted of 45 participants (37 female), aged between 18 and 28 years. All participants can be considered advanced learners of Dutch, given the fact that they were currently taking university degrees taught in Dutch, or had graduated from such a degree in recent years. Most participants had initially learned Dutch through an intensive five-week summer programme before starting their degree, of course in addition to mere exposure through immersion by living and/or studying in the Netherlands. All participants also reported knowledge of English, and some reported knowledge of further languages, mostly French and Spanish, although most participants indicated they rarely used these additional languages. None of the participants reported knowledge of Germanic languages other than Dutch, German and English.

A power analysis was not conducted because effect size estimates were not available in advance of this study: At this point in time, the De Vos, Schriefers, and Lemhöfer (2018) study had not yet been conducted, and to our knowledge there were no other L2 word learning studies where the participants were unaware of the study's aims to the same degree. Rather, we recruited as many participants as possible, although it was challenging to specifically target an immigrant population without appealing to their immigrant status or native language (which was needed to keep the participants unaware of the goal of the experiment).

Two thirds of the participants were assigned to the experimental group and one third to the control group.

This ratio was chosen because some of the research questions involved manipulations within the experimental group only. We started testing participants in the experimental group. The decision to include a control group was only made when the experiment had already been running for a while. Therefore, we then tested a number of participants in the control group to reach the desired ratio between the two groups. Subsequently, we alternated between testing participants in one group or the other.

Table 1 provides a comparison of the participants in the two groups on a number of dimensions that are known to affect L2 vocabulary learning. We used Welch t-tests when the data in both groups were normally distributed (as shown by a Shapiro-Wilk test), and Wilcoxon rank-sum tests otherwise. No significant differences between the participants in the two groups were found (all $ps \geq .32$). This shows that there were no systematic differences between the two groups with respect to dimensions that can be assumed to be relevant to vocabulary learning.

Materials

Target and filler words
Each participant was exposed to a total of 80 easy filler words and 24 to-be-learned target words (12 cognates and 12 non-cognates). These words were equally divided over four blocks, each block containing 20 filler words and six target words. The four blocks corresponded to four semantic categories ("children", "clothing", "household" and "tools"). We chose to present the items in semantic categories to make our price judgment cover story more credible; the participants may have been surprised if we had asked them to compare two completely unrelated objects. The specific categories were chosen because they contain many objects that are easy to recognise but often difficult to name in an L2, for example a whisk. Such items were potential target items. Potential fillers were items that were both easy to recognise and easy to name, even for L2 speakers, for example a glass.

We created the item pool by brainstorming and by looking through item lists of existing vocabulary studies. Group membership (for example, a whisk belonging in the household category) was decided intuitively. The "children" category contained objects or entities that children encounter on a regular basis, for example different toys, pets, and fruits. We did not consider it necessary to conduct a rating study of group membership since the categories were only used for the sake of the cover story, and all 24 target items would later be analysed together. As it turned out, during the experiment none of the participants commented on the group membership of the items.

After we had selected 250 potential target and filler items, as well as accompanying colour pictures which we had found on the internet, we pre-tested the total item set on 12 native speakers of German (L2 speakers of Dutch, not the participants in this study) and 12 native speakers of Dutch, in written online surveys. They were asked to provide the name of all the pictures in Dutch. On the basis of the names they wrote down, we selected the "best" 10 cognate target items and 10 non-cognate target items in every semantic category. "Good" target items were difficult to name for the German native speakers in the survey, while at the same time they evoked correct and stable names from the Dutch native speakers. In addition, the best 25 filler items were selected for each category. "Good" fillers received correct and consistent names from both German and Dutch native speakers. Cognate status was not controlled in fillers. Thus, the final item pool consisted of 40 cognate targets, 40 non-cognate, and 100 fillers. An example of a cognate target word is *schort* (German: *Schürze*, English: *apron*), an example of a non-cognate target word is *kwast* (German: *Pinsel*, English: *brush*). An example of a filler is *book* (German: *Buch*, English: *book*). A list of all the items can be found in the online supplementary materials that accompany this article on the *Language, Cognition and Neuroscience* website. As mentioned in the Introduction, the items (both targets and fillers) were selected on an individual basis for each participant. This means that from the final item

Table 1. Mean scores and standard deviations (between parentheses) on participant descriptives in the two conditions.

	Experimental n = 30	Control n = 15	Test statistics
Age	22.53 (2.47)	22.53 (2.50)	W = 228.5, p = .94
Years of learning Dutch	2.69 (1.78)	2.74 (1.96)	W = 230.5, p = .90
Self-rated proficiency*	3.07 (0.74)	3.27 (0.59)	W = 193, p = .41
Amount of daily exposure to Dutch*	3.07 (0.79)	3.29 (0.84)	W = 183.5, p = .32
Number of other languages known	2.33 (0.76)	2.47 (0.74)	W = 202.5, p = .56
Dutch vocabulary (LexTALE)	69.67 (7.75)	68.42 (8.27)	t(26.53) = 0.49, p = .63
Phonological working memory	80.17 (7.56)	81.71 (6.70)	t(28.53) = −0.68, p = .50

Note: For a description of the measurements, see Measures of individual differences. Variables marked with an asterisk were self-rated on a 1–5 Likert scale.

pool, a different subset was extracted for each participant. This will be discussed in more detail in the Procedure section.

The participants learned cognate words in two semantic categories and non-cognate words in the other two categories. Which semantic category was paired with which cognate status was counterbalanced across participants. The cognate and non-cognate items in each category were matched on several dimensions using the *Match* computer programme (Van Casteren & Davis, 2007). These dimensions, known to affect L2 word learning or processing, were word length (in phonemes) (e.g. Ellis & Beaton, 1993; Hulme, Maughan, & Brown, 1991) and L1 word frequency (e.g. De Groot, 2006; Lotto & De Groot, 1998). We also matched on compound status. Concreteness (De Groot, 2006; De Groot & Keijzer, 2000) was accounted for by only selecting depictable objects at the basic level of cognitive categorisation (Rosch, 1978). For example, we preferred a picture of a prototypical house cat over that of a special breed.

Measures of individual differences

The first five measures in Table 1 were obtained through a questionnaire. Self-rated Dutch proficiency was judged on a 1–5 scale (1 = *very bad*, 5 = *very good*). Self-rated exposure to Dutch was calculated as the mean of three other measures, all judged on a 1–5 scale (1 = *very rarely*, 5 = *very often*): How often do you read Dutch, how often do you speak Dutch, and how often do you watch Dutch television or listen to Dutch radio.

Phonological working memory in Dutch was measured through a non-word repetition task. The stimuli were taken from De Bree (2007), who had developed them for children at risk of dyslexia. We increased the stimuli's length to make them better suited to highly educated adult participants. The final stimuli set consisted of 16 non-words, ranging from three to six syllables. All the stimuli followed Dutch phonotactics, but neither the non-words nor their constituent syllables were existing Dutch lexical items. The stimuli can be found in the online supplementary materials.

Finally, Dutch vocabulary size was measured through the LexTALE vocabulary test (www.lextale.com; for the publication and validation of the English version, see Lemhöfer & Broersma, 2012).

Procedure

The participants were tested individually in a quiet lab. Before starting the experiment, they signed an informed consent form. They also consented to being audio-recorded during those tasks in which they would have to speak.

Sorting task (the pre-test)

The experimenter (a female native speaker of Dutch and the first author of this article) told the participants that the study was about making price judgments and that this would involve two tasks. In the first task (the sorting task), the participants would sort a pile of printed object pictures according to their estimated price. It was stressed that this ranking was subjective and there were no wrong answers, but that it was important that they remember their ranking for the second task. In that second, dialogue-like task (the price comparison task), they would see two object pictures in each trial and have to indicate which object was the cheaper one, consistent with their own ranking.

The sorting task acted as the secret pre-test of the participants' pre-existing active word knowledge. It was done by category and took approximately 30 minutes. After the participants finished sorting the 35 cards per category (10 potential target items and 25 potential fillers), they were told that they would now have the opportunity to consolidate their ranking once more by telling the experimenter out loud how they had sorted their cards. If they did not know an object's name in Dutch, they should describe it in Dutch with other words. For example, for a bib someone could say: "the thing babies wear when they eat". The experimenter sat behind a computer monitor and pretended to be coding the ranking, but was in fact coding whether or not the participant knew the object's name. In this way, we had a pre-test informing the experimenter which specific words a participant could produce in Dutch.

Selecting the target and filler items

After all four categories were pre-tested, the participant took a short break, while the experimenter prepared the price comparison task, in which the participants could learn the object names and would be tested on them. The experimenter ran the experimental software that selected, per category, six (actively) unknown target items out of the 10 pre-tested potential target items, and 20 (actively) known filler items out of the 25 pre-tested potential filler items. If less than six unknown target items were available for a category, the participant still finished the experiment, but was excluded from the analysis (later into data collection, we immediately aborted the experiment at this stage, although the participant would still get paid). This was the case for 15 participants. If less than 20 known fillers were available for a category, other known fillers would appear slightly more often. The lower limit for participation was set at 15 known fillers per category, and all participants reached this criterion.

Price comparison task (the learning phase)

After the selection of targets and fillers was completed, the participant and the experimenter continued to the price comparison task, which took the form of a dialogue between the experimenter and the participant. In this way, we approximated an L2 conversation in the lab. The participants later often reported that they thought the interaction with the experimenter was meant to influence their perception of prices. The price comparison task also took approximately 30 minutes. The participant and experimenter sat behind opposite computer monitors and keyboards, and could not see the other person's monitor. The price comparison task consisted of 82 trials per semantic category, 328 in total, presented with PsychoPy (Peirce, 2009). The order in which the four categories were presented was the same as during the sorting task, and was counterbalanced across participants. On each trial, two object pictures appeared next to each other on the screen, both filling an imaginary rectangle of 15×15 cm. A trial either consisted of a target item and a filler item, or of two filler items.

The experimenter and the participant took turns in stating out loud a judgment concerning the price of the two objects on the screen, for example: "A bib is cheaper than a t-shirt". The participants had to make this statement based on their own insight in object prices, and were told to try to adhere to the ranking they had made during the sorting task. After the participant's statement, the experimenter pressed the button (pretending to make a price judgment, but in fact coding whether the participant had correctly produced the target word). The participants had been instructed to try using Dutch names for the objects, but could again resort to Dutch descriptions if they did not know an object's name. The experimenter's statements were scripted and were always reasonable, although not always in accordance with the ranking the participant had made during the sorting task. After the experimenter's statement, the participant's task was to press a button to express agreement or disagreement with the experimenter's price judgment. There was no time limit for these button presses, and they were not analysed since we were not actually interested in the participants' perception of object prices. The next trial appeared immediately after the button press. Between the categories the participants could take a short break.

For the participants in the experimental group, all target items were named by the experimenter (in her trials) twice before appearing in the participant's trial for the first time. In other words, the participants had twice been exposed to a target object's name before being first tested on it. The test took place either three or seven trials after the last exposure. This represents the predictor Lag. Which item was associated with which lag was counterbalanced across the participants. After one "round" of two exposures and one test was finished for all six target items, the second round began. All target items again were produced twice by the experimenter, and then once by the participant (after three or seven trials). This was the second testing moment, allowing us to examine the predictor Exposure frequency. Within a round, the inter-stimulus interval between the two exposures to a target word was always fixed at five trials. Between the rounds, this interval was not fixed.

For the participants in the control group, none of the target items were named by the experimenter. Instead, the experimenter's trials only contained fillers. This means that the predictors Exposure frequency and Lag were essentially meaningless for the participants in this group. Please recall that the control group was included to investigate whether participants might have, or develop, potential productive knowledge of target items which they did not display in the pre-test. Therefore, the control participants also had to produce the target items in their trials, and these target items had been selected individually based on the participants' pre-existing knowledge.

Debriefing and additional tests

After the price comparison task was finished, the participants were asked what they thought the experiment was about and were subsequently told its true aims. Then, they filled in the personal and language background questionnaire, and took the phonological working memory task and the LexTALE vocabulary test.

First post-test

The participants were then presented with an unannounced post-test (this was the third test of each item). This post-test took place approximately 20 minutes after the end of the price comparison task and was an explicit picture-naming task. The participants saw, one by one, pictures of all target and some filler objects on the screen and were asked to name them. The experimenter then provided them with the correct name. Finally, the participants had to indicate whether they were familiar with the 12 cognates' German translations. If this was not the case for one or more words, these words would be excluded from the analysis. The reasoning is that if participants did not know an L1 word form, then the related L2 target words could not benefit from the hypothesised cognate advantage.

Second post-test

Six months after their participation, the participants in the experimental group were contacted by e-mail to

ask if they were willing to return to the lab to once more name the objects from the experiment. They did not know they would be invited for this follow-up, which comprised the fourth test of the target items. Eighteen of the participants in the experimental group returned (two of them on Skype) and performed the explicit picture naming test again, which was the same as the 20-minute delayed post-test. After trying to name each target item, they were provided with its correct name and were asked whether they had encountered this word in the last six months. Because the results of the control group did not show any change during the first three testing moments (see Results), for logistical reasons the participants in the control group were not invited to come back for the follow-up test.

Analysis

Measures of individual differences

The measures of individual differences were used to describe and compare the participants in the experimental and control group (see Table 1). We did not have specific hypotheses for the relationship between these measures and L2 word learning in a non-learning-centred setting such as the current one. Since we were wary of overfitting our model, we left these measures out of the main statistical analysis. However, explorative correlations are reported in Appendix 6 in the online supplementary materials.[1]

Data preparation

The following responses to target words were excluded from the data set:

- Words for which the participants had displayed partial knowledge in the pre-test (2.8% of the total data set), for example when saying *wafelding* (literally in English: *waffle thing*) instead of *wafelijzer* (English: *waffle iron*).
- Words for which the participants had used a correct synonym in the pre-test, which made it impossible to see whether or not they knew the name that we used throughout the experiment (0.3% of the remaining data set), for example, using *haarspeld* or *haarclip* (in English comparable to *hair pin* and *hair clip*) for the target word *speldje* (meaning *hair pin/hair clip*).
- Cognate words for which the participants later indicated they did not know the German name (3.9% of the remaining data set).
- From the analysis of the second post-test, those words were excluded for which the participants indicated through self-report that they had encountered them in the six months following the experiment. For

these words, we could not know whether any potential knowledge would be due to our experiment, or to other forms of exposure (0.5% of the remaining data set).

Overall, 7.2% of the data points (i.e. target word productions) were removed from the total data set. This left 3407 data points, from 45 participants, for analysis.

Scoring

The participants sometimes produced target word utterances that were neither correct, nor fully incorrect. An example would be a participant saying *gorde* rather than (correct) *garde* (English: *whisk*). To capture this nuance, we scored the data at the phoneme rather than the word level. Phonemes were scored as incorrect if they had been deleted, inserted or substituted by another phoneme (see Levenshtein, 1966). The *gorde* example thus would be scored as the vector (4, 1), indicating four correct phonemes and one incorrect (substituted) phoneme. Of course, a correct response in this case would have been scored (5, 0), and an incorrect response (0, 5). Responses that were obviously wrong, such as *parfum* (English: *perfume*) for the picture of the whisk were always scored as fully incorrect, even if one or more phonemes would incidentally overlap (here: ɑr). For descriptive statistics, we converted the ratios of correct and incorrect phonemes to percentages (80% correct in the above example). For a more elaborate description of the scoring method, see De Vos, Schriefers, and Lemhöfer (2018).

Modelling

The data were analysed with two generalised linear mixed-effects models of the binomial family, with the logit link function. The binomial distribution describes the probability of achieving a particular number of "successes" in a sequence of N independent trials. In the above *gorde* example, we would model the probability of producing four out of five phonemes correctly. The vector (4,1), representing (Number of correct phonemes, Number of incorrect phonemes), would in this case be the dependent variable. Crawley (2007, pp. 569–570) discusses four reasons why such vectors are preferred to percentages (here: 80%) as the dependent variable for the statistical analysis of proportion data. These include the fact that proportions are bounded between 0 and 1, that the variance is non-constant, and that the errors are non-normally distributed.

We created one statistical model to focus on the participants' word learning (i.e. Research questions 1 and 2), and a second model to focus on the participants' retention of the words they had learned in the experiment (i.e.

Research question 3). These models are referred to as the *learning model* and the *retention model* respectively. In the learning model we modelled the scores the participants had obtained on the two testing moments in the price comparison task, when they had been prompted to produce the target words after two and four exposures. In the retention model we modelled the scores the participants had obtained in the two explicit post-tests, and compared these scores to the participants' last scores obtained during the price comparison task (i.e. after four exposures), when their newly acquired word knowledge was at its peak.

Included as fixed effects in the learning model were the main effects of Group (experimental versus control), Cognate status (cognate versus non-cognate), Exposure frequency (two versus four exposures), and Lag (three versus seven trials). Following our hypotheses, we investigated the main effects of Cognate status, Exposure frequency and Lag in the experimental group only (please recall that Exposure frequency and Lag were meaningless in the control group, since the control participants did not receive input on the target items). We also investigated the interaction of these predictors with Group. If such an interaction is significant, this shows us that it was the exposure to input underlying any potential effects of the predictors, and that these effects did not just arise as the result of guessing and/or repeated testing. In Appendix 3 in the online supplementary materials we also report additional models, with which we explored other potential interactions between the predictors. We will call these models the *explorative models*. They are meant to identify potentially interesting patterns in the data that can be further examined in future research. The models reported in this text are the *hypothesis-based models*.

In the retention model, we included the main effects of Cognate status, Retention interval and Lag as fixed effects. Group was left out; this time, we only considered the scores of the participants in the experimental group. The participants in the control group were not included in the retention analysis because they had had no opportunity to learn the target words. Therefore, no retention was possible either.

We did not have any hypotheses regarding the random-effects structure for either the learning or the retention model. To establish an appropriate random-effects structure, we started with a model with only the above mentioned fixed effects, and random intercepts for participants and for words. These intercepts represent the random variability in participants' word learning abilities, and the random variability in learnability between words. Then, for the learning and retention models separately, we systematically assessed potential random slopes one by one. Each time the model converged (i.e. if it could be computed), we checked with a likelihood ratio test whether the model with the new random slope was a significantly better fit to the data than a model without this random slope. We also checked whether this coincided with a decrease in the Akaike Information Criterion (AIC; Akaike, 1974), and whether the new random slope could be supported by the data, in other words, whether the model was not overparameterised (following Bates, Kliegl, Vasishth, & Baayen, 2015). If all these criteria were met, we included the random slope in the model and assessed the next random slope. If not all the criteria were met, we removed the random slope from the model, added the next random slope, and compared this model to the last model that had met all the criteria. This process was continued until the random slopes of all main effects and their interactions had been explored (except that we did not explore higher-order interactions if the random slopes of lower-order effects did not meet the criteria). These model comparisons are reported in Appendix 2 in the online supplementary materials; the final models are presented in the results section.

All models were computed using R's "lme4" package (Bates, Mächler, Bolker, & Walker, 2015; version 1.1–12) in R (R Core Team, 2018). Because of convergence problems with the default optimisation settings, we used the "bobyqa" optimiser (Bound Optimization BY Quadratic Approximation; Powell, 2009). The maximum number of iterations for the optimiser was set to 100,000. Alpha was set at .05.

Results

Hiding the goal of the study

Out of the 61 participants tested, only one correctly guessed that the study had been about word learning. She was excluded from the analysis. The other participants believed that the study had been about (consistency in) making price judgments, and had not been aware that the study was specifically targeted at German native speakers and concerned word learning.

Descriptive statistics

The learning scores are depicted graphically in Figure 1. Pre-test scores were at zero for everyone, since our software had selected unknown target words for each participant on an individual basis.

In Tables 2 and 3, descriptive statistics are shown per predictor (split by Group), for learning and retention separately. Table A in online Appendix 1 contains
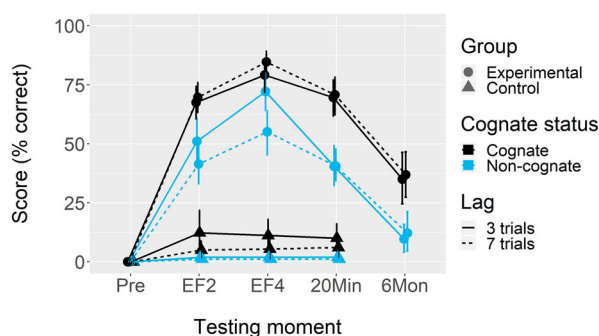
Figure 1. Mean scores across the four testing moments (EF = Exposure frequency). Error bars represent 95% confidence intervals based on a bootstrap.

descriptives for all subcombinations of predictor levels as well. As explained in the Scoring section, in both the figure and the tables the dependent variable is the average percentage of correctly produced phonemes per target word utterance.

As can be seen from these results, there is clear effect of Group: Vocabulary scores were much higher for the participants who were exposed to input (i.e. the experimental group). It is interesting to see, however, that despite the pre-test and the following individualised item selection, the average score in the control group is not zero, especially for cognates. An effect of Cognate status is also seen in the experimental group. It is only in the experimental group that an effect of Exposure frequency is visible, which is unsurprising given that Exposure frequency was a meaningless predictor in the control group (there was no exposure to the target items at all). Finally, in Figure 1, there seems

to be an interaction between Cognate status and Lag in the experimental group: Participants seemed to perform better on non-cognates when being tested after a lag of seven trials as compared to three trials. However, this was not the case for cognates, where if anything the effect was reversed. We had no hypothesis about the presence of such an interaction, and explored it further in Appendix 3 in the online supplementary materials. Online Appendices 6 and 7 contain additional analyses at the participant and item level.

Model comparisons

Online Appendix 2 contains the results of the model comparisons we performed for finding the best-fitting model for the data from the learning phase. The final learning model was: (Number of correct phonemes, Number of incorrect phonemes) ~ 1 + Group * Cognate status + Group * Exposure frequency + Group * Lag + (1 + Cognate status * Exposure frequency * Lag | Participant) + (1 + Group * Lag + Exposure frequency * Lag | Word). In this notation based on the R programming language, the dependent variable on the left of the "~" is modelled from the fixed and random effects positioned on the right of the "~", "1" represents an intercept, "|" represents random effects, and "*" represents an interaction including all lower-order effects. For example, Group * Cognate status represents the main effects of Group and Cognate status, as well as their interaction.

The model comparisons we performed for finding the best retention model are also shown in online Appendix 2. The final retention model was: (Number of correct phonemes, Number of incorrect phonemes) ~ 1 + Cognate

Table 2. Percentage of correctly produced phonemes per target word during the price comparison task (i.e. the learning phase).

		Experimental group			Control group		
		Mean	SD	95% CI	Mean	SD	95% CI
Cognate status	Cognate	75.26	12.19	70.71–79.82	8.47	8.20	3.93–13.02
	Non-cognate	54.93	18.43	48.05–61.82	1.50	4.34	−0.91–3.91
Exposure frequency	2 times	57.43	14.87	51.88–62.98	5.08	5.64	1.96–8.20
	4 times	72.77	14.07	67.51–78.02	4.89	4.05	2.65–7.13
Lag	3 trials	67.48	16.89	61.17–73.79	6.82	7.93	2.43–11.21
	7 trials	62.72	14.57	57.28–68.16	3.15	4.36	0.74–5.57
Total		65.10	13.60	60.02–70.18	4.99	4.73	2.37–7.60

Table 3. Percentage of correctly produced phonemes per target word in the two post-tests (i.e. the retention phase).

		Experimental group			Control group		
		Mean	SD	95% CI	Mean	SD	95% CI
Cognate status	Cognate	59.06	14.97	53.47–64.65	8.04	6.91	4.21–11.87
	Non-cognate	31.27	17.24	24.83–37.71	1.50	4.34	−0.91–3.91
Retention interval	20 min	55.36	14.83	49.82–60.90	4.77	4.05	2.53–7.01
	6 months	23.43	12.68	17.13–29.73	N/A	N/A	N/A
Lag	3 trials	44.00	14.04	38.76–49.24	5.95	6.65	2.26–9.63
	7 trials	46.32	17.58	39.76–52.89	3.59	4.86	0.90–6.28
Total		55.36	14.83	49.82–60.90	4.77	4.05	2.53–7.01

status + Retention interval + Lag + (1 + Cognate status * Retention interval + Lag | Participant) + (1 + Retention interval * Lag | Word). Below, we will describe how the final models' fit to the data was evaluated.

Inferential statistics

We will now evaluate the statistical evidence for the effects that we previously described based on visual inspection. The learning phase (i.e. the price comparison task) and the retention phase (i.e. the two explicit post-tests) were analysed separately. Table 4 shows the model estimates and test statistics for the learning phase, in which the participants were exposed to correct input and tested both after two and four exposures to the target words. Table 5, presented below, contains the long-term retention results. We will begin with explaining how these tables should be interpreted, and then turn to the actual outcomes.

Interpretation of model estimates
Please note that all effects should be interpreted relative to the intercept, which represents a specific combination of predictor levels (see the note under Tables 4 and 5). For example, in Table 4, we can see that there is a positive effect of having four exposures ("EF = 4 times"), as compared to the level of Exposure frequency represented by the intercept (i.e. two exposures).

It is also important to understand that the main effect of Group ("Group = Control") specifically applies to cognate words tested after two exposures, presented

with a lag of three trials. This is because the interactions between Group and the three other predictors were included in the model as fixed effects (the last three of the fixed effects in Table 4). In the hypothesis-based learning model reported here, we did not include any fixed-effect interactions that did not include Group (in the explorative model, reported in online Appendix 3, these other interactions were included). For this reason, the interpretation of the main effect of Group is different from the interpretation of the main effects of Cognate status, Exposure frequency and Lag. Each of these three main effects applies to the experimental group only, and has been calculated by collapsing over the levels of the other predictors. For example, the main effect of Cognate status for the experimental group has been calculated using the data of both exposure frequencies and both lags.

Effect sizes are expressed as odds ratios (ORs). The OR tells us how the odds of correctly producing a phoneme change for one predictor level as compared to the level of that predictor that is represented by the intercept. ORs that are much higher than 1, or that are very close to zero, indicate large effects. The exact interpretation of ORs, as well as the interpretation of logit estimates, is explained in more detail in Appendix 4 in the online supplementary materials.

Outcomes of the learning phase
As can be seen from the main Group effect in Table 4, the participants in the experimental group significantly out-performed the participants in the control group. This

Table 4. Outcomes of the learning model.

Fixed effects	Logit	Odds ratio	SE	z	p
(Intercept)	2.80	16.42	0.76	3.71	**<.001**
G = Control	−11.96	<0.001	2.24	−5.33	**<.001**
CS = Non-cognate	−3.25	0.04	0.75	−4.35	**<.001**
EF = 4 times	1.72	5.60	0.28	6.13	**<.001**
L = 7 trials	−0.74	0.48	0.60	−1.23	.22
G = Control: CS = Non-cognate	−2.86	0.06	2.55	−1.12	.26
G = Control: EF = 4 times	−2.26	0.10	0.71	−3.17	**.002**
G = Control: L = 7 trials	−3.36	0.03	3.37	−1.00	.32

Random effects	Variance	SD			
Participant (Intercept)	5.55	2.36			
CS = Non-cognate	4.61	2.15			
EF = 4 times	1.79	1.34			
L = 7 trials	5.46	2.34			
CS = Non-cognate: EF = 4 times	2.53	1.59			
CS = Non-cognate: L = 7 trials	16.11	4.01			
EF = 4 times: L = 7 trials	5.94	2.44			
CS = Non-cognate: EF = 4 times: L = 7 trials	5.85	2.42			
Word (Intercept)	11.47	3.39			
G = Control	48.06	6.93			
EF = 4 times	3.80	1.95			
L = 7 trials	8.01	2.83			
G = Control: L = 7 trials	133.27	11.54			
EF = 4 times: L = 7 trials	6.37	2.52			

Note: The intercept represents the following combination of variable levels: G [Group] = Experimental, CS [Cognate status] = Cognate, EF [Exposure frequency] = 2 times, L [Lag] = 3 trials. Colons (:) represent interactions but not lower-order effects, equal signs (=) signal the level of a categorical variable. Significant p-values are printed in bold.

Table 5. Outcomes of the retention model.

Fixed effects	Logit	Odds ratio	SE	z	p
(Intercept)	3.79	44.13	0.56	6.73	**<.001**
CS = Non-cognate	−3.07	0.05	0.58	−5.25	**<.001**
RI = 20 min	−1.62	0.20	0.26	−6.15	**<.001**
RI = 6 months	−6.15	0.002	0.72	−8.49	**<.001**
L = 7 trials	−0.18	0.84	0.33	−0.54	.59
Random effects	Variance	SD			
Participant (Intercept)	2.31	1.52			
CS = Non-cognate	1.75	1.32			
RI = 20 min	0.64	0.80			
RI = 6 months	5.19	2.28			
L = 7 trials	0.91	0.95			
CS = Non-cognate: RI = 20 min	0.69	0.83			
CS = Non-cognate: RI = 6 months	21.67	4.66			
Word (Intercept)	13.79	3.71			
RI = 20 min	2.75	1.66			
RI = 6 months	23.14	4.81			
L = 7 trials	5.74	2.40			
RI = 20 min: L = 7 trials	3.91	1.98			
RI = 6 months: L = 7 trials	24.57	4.96			

Note: The intercept represents the following combination of variable levels: CS [Cognate status] = Cognate, RI [Retention interval] = 4 exposures (i.e. participants' scores after 20 minutes and six months are compared to their last score from the price comparison task), L [Lag] = 3 trials. Colons (:) represent interactions but not lower-order effects, equal signs (=) signal the level of a categorical variable. Significant *p*-values are printed in bold.

indicates that exposure to spoken L2 input in interaction can result in the acquisition of new vocabulary. The OR was very large. As explained above, this effect specifically applies to cognate words tested after two exposures, which were tested after a lag of three trials. However, the Group effect for non-cognates, and the Group effect after a seven-trial lag, were not significantly different from the Group effect for cognates after a three-trial lag ($p = .26$ and $p = .32$). The effect of Group did grow significantly more pronounced after four exposures as compared to two exposures ($p = .002$). Averaged over all other predictors, the experimental group learned about 1205% (or 13.05 times) more phonemes than the control group.

Having shown how Group interacts with the other predictors, we will now focus on the main effects of Cognate status, Exposure frequency and Lag in the experimental group only (in accordance with our hypotheses). Cognate status had a significant and large effect: Participants in the experimental group learned 37% more phonemes in cognate words as compared to non-cognate words. With regard to Exposure frequency, the experimental participants had learned 27% more phonemes after four as compared to two exposures. This effect also was significant, with a medium-to-large effect size. No main effect of Lag could be detected in the experimental group, and the effect size was negligible. The explorative learning model reported in Appendix 3 showed that the interaction between Group, Cognate status and Lag during the learning phase that seemed visible in Figure 1 did not reach significance during the learning phase.

Long-term outcomes

To investigate long-term retention, we turn to Table 5.

At the time of the first post-test, 20 minutes after the end of the price comparison task, word knowledge in the experimental group had significantly dropped, as compared to scores during the price comparison task after four exposures. The participants remembered 24% fewer phonemes, and the effect size of this decay was medium-to-large. At the time of the second post-test, six months after the price comparison task, the participants remembered 68% fewer phonemes as compared to when tested directly after four exposures. This contrast was highly significant, with a very large effect size. Relevelling of the model by making the second post-test the intercept showed that in comparison to the first post-test, scores had declined by 58% ($β = −4.70$, $OR = 0.01$, $z = −7.67$, $p < .001$); the effect size was very large. Yet, the intercept in this model was still significant ($β = −2.50$, $z = −3.01$, $p = .002$). This tells us that even after six months, the participants' scores were still significantly above zero.

The explorative retention model presented in online Appendix 3 showed that, between the last testing moment in the price comparison task and the second post-test six months later, cognates were forgotten at significantly different rates from non-cognates (there was more decay for non-cognates). Between the last testing moment in the price comparison task and the first post-test 20 minutes later, there was also a significant interaction involving both Cognate status and Lag: For non-cognates, words that had originally been tested after a lag of three trials were forgotten at a

higher rate than words that had originally been tested after a lag of seven trials. For cognate words, the effect was reversed, and less strong.

Models' goodness of fit

In this section, we summarise the outcomes of the evaluation of our models' fit to the data, which is reported in detail in online Appendix 5. While we found that the errors in our learning model were not uniformly distributed, our model fitted the data better than an alternative model (with a different random-effects structure) that had a more uniform distribution of errors. The model estimates and significance values were very similar for these two models, which shows that we can be confident in the outcomes of our learning analysis. In addition, from Table 4 it can be seen that none of the variance components in the random-effects structure were at zero. Furthermore (not presented in this text), none of the correlations between the random effects were at (−)1 or close to (−)1, the highest one being −.88 but most correlations being much lower. Both of these observations suggest that the model was not overparameterised (Bates, Kliegl et al., 2015, p. 7).

With regard to the retention model, the distribution of the residuals seemed to be acceptable, but the model's predictions tended to overestimate the very low scores. A likely explanation for this finding is the absence of low scores in our data set, whereas our model was set up to make continuous predictions (also for low scores). However, as pointed out in Appendix 5, in our analyses we focused on contrasts, and not so much on absolute scores. Therefore, we did not consider the model's bias in the low domain (i.e. scores between 0 and ±0.10) to be a relevant concern.

Discussion

In this study, we investigated interactive L2 word learning in immersed learners who were unaware of taking part in a language learning study. We introduced a novel and well-controlled experimental setting in which the predictors Cognate status, Exposure frequency and Lag were manipulated. Twenty minutes and six months after the experiment, it was measured how well the participants had retained the words from the experiment. As described in the Results section, all but one of the participants (who was excluded from the analysis) remained unaware of the study's language learning aspect until the experimenter debriefed them. With this, we clearly reached our goal of creating a setting to approximate real-life L2 learning in the lab, although we should point out that the participants' learning behaviour most likely was intentional rather than incidental, as will be explained in the next section. This does not mean that the learning we observed was not naturalistic, since language learning in real-life settings can also be intentional. However, it does mean that the learning we observed probably concerns situations in which L2 learners try to learn a new word, for example when encountering an object and asking their conversational partner what it is called.

High absolute gains

Our first research question was what L2 word learning in an interactive immersion setting looks like in the context we described earlier. Overall, we conclude that exposure to spoken L2 input in a dialogue-like setting can result in large vocabulary gains. This was seen from the experimental group (which received target word input) significantly outperforming the control group (which was only exposed to filler words), with a very large effect size. In fact, overall performance on the target words during the learning phase was 1205% (or 13.05 times) better for the experimental group than for the control group. Several possible explanations for the magnitude of this effect are given below.

First, it was relatively easy for the participants to establish form-meaning links between the target words and the objects they represented. Each object was named by the experimenter while the participants looked at the corresponding picture. In such a setting, it is likely that fewer exposures are needed as compared to settings where learners need to infer the meaning from a purely communicative context.

Second, although each participant was exposed to a selection of target items that he or she had been unable to name during the pre-test, it is possible that the participants already had receptive knowledge of some of the target words. This may also have contributed to their high overall learning scores. Still, this would be no different in naturalistic learning situations. The contribution of pre-existing passive knowledge to L2 word learning is explored by De Vos, Schriefers, and Lemhöfer (2018), who found that such knowledge was beneficial for participants who noticed holes in their vocabulary.

Third, while they were not instructed to learn words, a few trials into the price comparison task the participants may have realised that they would have to name all objects. Thus, they may have tried to learn from the experimenter's utterances in anticipation of their upcoming turns, especially if they wanted to make a good impression on the experimenter, who interacted with them throughout the experiment. As a result, they probably developed some intention to learn words, and were perhaps internally preparing for the production moments.

This latter explanation is supported by the findings of De Vos, Schriefers, and Lemhöfer (2018), whose design was less interactive than the design of the current study. Their participants did not speak during the price comparison task, but only listened to input (four exposures per target word, non-cognates only). This means that these participants probably were not anticipating to produce the target words in front of the experimenter. They only achieved post-test scores of around 28% after 15 minutes, while in the current study the post-test scores for non-cognates were around 41% after 20 minutes. Thus, the anticipation of their upcoming turn in our dialogue-like setting seems to have increased the participants' motivation for learning.

There is an additional difference between the two studies that can also explain why the scores in the current study were higher than in De Vos, Schriefers, and Lemhöfer (2018): Our participants could benefit from retrieval practice during the learning phase. At the time the post-tests took place, the participants had already been tested on the target words twice before. It has been shown that trying to retrieve newly studied words from memory facilitates their retention over time (Barcroft, 2007).

Finally, De Vos, Schriefers, and Lemhöfer (2018) showed that noticing vocabulary holes benefits word learning as well. Our pre-test induced the noticing of vocabulary holes: The participants were asked to name the target words out loud, but generally were not able to do so. At these moments, they noticed the holes in their vocabulary. Then, in the price comparison task, they were exposed to input that contained the vocabulary they had just noticed missing. This can also explain why, in an absolute sense, the learning scores in the current study were quite high.

The above observations, specifically the supposed intentional learning behaviour of our participants and the fact that L2 word forms were presented together with object pictures, give rise to the idea that the kind of learning exhibited by our participants may have been comparable with paired-associated learning. In the context of word learning, this is a form of learning where L2 words are presented together with their L1 translations or a picture. Paired-associate learning is a form of intentional learning, and is typically shown to be very effective (e.g. Hulstijn, Hollander, & Greidanus, 1996; Mondria, 2003). In fact, our learning rates for cognates (75%) and non-cognates (55%) were close to those reported in De Groot and Keijzer (2000), who let their participants learn cognates and non-cognates in a paired-associated paradigm. After two exposures, they found learning rates of 70% for cognates and 44% for non-cognates.

Predictors of L2 word learning

The hypothesis-based model showed that the participants in the experimental group acquired cognates at significantly higher rates than non-cognates (with a large effect size). The cognate advantage is in line with the literature (e.g. Lotto & De Groot, 1998). However, the cognate effect in the control group was not significantly different from that in the experimental group, suggesting that the control group also benefited from a cognate advantage.

The fact that we coded correctness on the phoneme rather than the word level is relevant for explaining this last finding. Remember that our dependent variable was based on the number of phonemes that were produced correctly and incorrectly. In other words, participants could still obtain a high score when they produced partially correct versions of many words, even if they did not produce any words fully correctly. In the raw data (not presented here, but available online; see Data availability statement), it can be seen that the participants in the control group on average produced partially correct responses for 11% of cognates, but only for 1% of non-cognates. In contrast, the percentage of fully correct responses was the same across cognate status: 1% for both cognates and non-cognates. Thus, it seems that the cognate effect in the control group can be explained by the participants making educated guesses based on their L1 knowledge, which resulted in a partially correct response.

In the experimental group, partially correct responses were produced as well (16% of cognate responses, and 11% of non-cognate responses). However, in this group a partially correct response did not necessarily mean that the participant was making an educated guess: A partially correct response could also represent an incomplete representation of the word form in memory, as the result of previous exposure. Even if we assume that a partially correct response was always due to guessing, and a fully correct response was due to actual knowledge of the word form, then guessing could not explain the cognate effect in the experimental group. The reason for this is that the percentage of fully correct responses was also higher for cognates (63%) than for non-cognates (43%). In fact, the ratio is almost exactly the same: $16/11 \approx 63/43$.

The question is still open as to why the participants in the control group only started guessing during the price comparison task, and not already during the sorting task (i.e. the pre-test). We know they behaved differently during the two tasks because all of the target words in the price comparison task were words that the participants had not shown any productive knowledge of

during the sorting task (this is why the pre-test scores are at zero in Figure 1). It cannot be due to the presence of the experimenter, since she was present during both tasks. Perhaps the price comparison task felt slightly more formal to the participants, as the participant and the experimenter always alternated in naming the two objects, and the participants therefore would have felt a higher need to make guesses.

Still, even if it is not entirely clear why the price comparison task made the participants more inclined to guess, it is likely that this effect was the same for the participants in the experimental and control group. The fact that the experimental group achieved learning scores so much higher than those of the control group indicates that it was the exposure to the target words causing the effect, and not just guessing or repeated testing. Finally, the control group not scoring at zero is in line with the meta-regression by De Vos, Schriefers, Nivard, et al. (2018), which also showed that effect sizes in studies with a true control group that is not exposed to input are significantly smaller. This shows the importance of including no-input control groups in L2 studies (especially when cognates are used as target items), which currently only seems to be done in a minority of studies.

Exposure frequency

The first two exposures (taken together) had a bigger impact on learning than the third and fourth exposure (taken together). This can be seen in Figure 1 from the learning gains being larger after two exposures as compared to four. Still, the participants produced significantly more correct phonemes after four exposures than after two exposures, which is unsurprising because this testing moment represents the cumulative effect of all exposures combined. Relatedly, it is easy to explain why the effect of Group was significantly stronger after four than after two exposures: Only the scores of the experimental participants kept rising between two and four exposures, while the scores of the participants in the control group remained constant throughout the experiment, as they were not exposed to the target words.

The finding that the first two exposures had relatively more impact than the following two exposures is one that is obtained in paired-associated word learning studies as well (e.g. De Groot & Keijzer, 2000). It also resembles the findings of Bisson et al. (2014). They operationalised and measured learning differently, but found an incidental learning effect of 6% after two exposures, and 7% after four exposures. The explanation mentioned above, about why relatively few exposures are needed to establish form-meaning links, is also given by Bisson et al.

(2014, p. 871) to explain their non-linear effect of exposure frequency. In addition, when the target words were presented for the first time, they may have attracted extra attention from the participants due to their novelty, and this effect may have worn off over time (Bisson et al., 2014, p. 872). In future studies, it would be interesting to measure word learning after each additional exposure (instead of pairs of two exposures), and perhaps to employ some online measurements to see whether earlier exposures indeed attract more attention from learners. As Bisson et al. (2014, p. 872) suggest, eye tracking may be a good candidate for this.

Our findings differ from Vidal's (2011) findings for learning from listening. Her participants watched a video recording of three academic lectures, and were tested on vocabulary afterwards. The frequency of occurrence of the target words was one, two, three, four, five or six times. The learning curve practically stayed flat between one and three exposures, then rose slightly between three and five exposures, and suddenly rose steeply at six exposures. Thus, there was no steep initial rise, followed by a more gradual rise later on, like in the current study. The explanation regarding form-meaning links could also apply here: The participants in Vidal (2011) might have needed more repetitions because they had to derive the meaning of the target words from context in the academic lecture.

Lag

No main effect of Lag was found in the experimental group, and its effect in the control group was not different from that in the experimental group. Perhaps the difference between the two lags was too small to evoke any effect. After all, the difference between a test either three or seven trials after exposure was only about 20 seconds.

However, we had also noticed that there seemed to be a deviant outcome in the data set: After four exposures, the participants in the experimental group scored atypically high on non-cognates when tested after a lag of seven trials (see Figure 1). Still, this interaction between Cognate status and Lag (in the experimental group) was not significantly different after four exposures as compared to two exposures (see online Appendix 3). In contrast, the interaction between Cognate status and Lag was significantly different after four exposures as compared to twenty minutes after the price comparison task. By then, the difference between non-cognates that had first been tested after three versus seven trials had disappeared. It seems that the significance of this interaction was carried by the deviant data point described above. We had no hypothesis about this data point, but

rather detected the significant interaction it was involved in when running an explorative model that included all possible interactions in the data set. We therefore draw no further conclusions from this finding. First, it should be replicated in hypothesis-based research.

Long-term retention

The third research question concerned the retention of the newly acquired words. Twenty minutes after the experiment, the scores of the experimental group had dropped approximately 24% as compared to their scores after four exposures. This was a significant and large decline. Six months later, the scores had declined about 68% relatively to their scores after four exposures, but were still significantly above zero. In calculating these proportions, words that the participants reported to have encountered in the six months following the experiment had already been excluded. Thus, considering that six months ago they had received input on the target words only four or five times (a fifth time in case of an incorrect answer during the first post-test, when the experimenter provided them with the correct answer), these outcomes are remarkable.

Relation to the immersion and incidental learning literature

At the beginning of this article, we briefly introduced the research domains of immersion and incidental learning. With its experimental approach to L2 learning in an immersion setting, the current study complements the existing, mostly non-experimental immersion literature. With regard to incidental learning, we mentioned that participants in incidental learning studies can generally deduce that a study is about L2 learning, even if they are not explicitly told so. The current study differs from this research tradition in keeping the participants unaware of the study's purpose throughout the learning phase.

Since awareness of the study's language learning aspect plays such a central role throughout this article, it would be interesting to investigate in future research what is the actual impact of such awareness on L2 word learning. The current study could be extended to investigate this question. For example, the same task could be repeated in an L2 classroom, which would likely induce the participants' suspicion regarding the study's language learning aspect. Alternatively, the study could still be conducted in a lab, but this time the participants' native language could be mentioned during recruitment (for example: "German native speakers needed for price judgment task"). Optionally, an extra

group could also be added in which participants are explicitly instructed to learn words, in order to study the effects of such instruction.

In addition to the (non-)awareness factor, the learning in the current study does not fully overlap with "typical" incidental learning in other aspects either. As explained above, our participants probably developed an intention to learn words and expected to be prompted to produce these words during the price comparison task. This means that the learning does not seem to have been incidental with regard to the first and second definitions of incidental learning as they were given in the Introduction. However, learners who engage in immersed L2 interaction might also develop the intention to learn words from their conversational partners from time to time, or plan to incorporate newly-learned words in their upcoming utterances. Thus, in this sense, the current study seems to be more representative of real-life L2 word learning in conversation than do typical studies on interactive incidental L2 word learning.

A methodological innovation, as compared to the existing literature, was that we used a new approach to item selection. Our experimental software selected the target and filler items on a by-participant basis by using the outcomes of the sorting task. This made it possible to work with words from a language that the participants already had been using in daily life (here: Dutch). While the participants often had different pre-existing knowledge of Dutch, our on-the-spot item selection ensured that they all learned an equal number of previously (productively) unknown words, and thus experienced a similar memory load.

Summary and conclusions

This study showed that participants who are unaware of taking part in an L2 word learning study can learn from interaction with a native speaker at high rates. Despite being unaware of the study's purpose, it is very well possible that the participants developed an intention to learn words, due to various aspects of our experimental procedure and design. This probably led the participants to make an effort to remember the target words they encountered, and means that our results are most representative of situations in which learners are consciously trying to learn a new word from spoken input.

The learning rates were dependent on exposure frequency: Four exposures led to more learning than two exposures, although relatively speaking, more learning happened in the first two as compared to the last two exposures. Cognate words were acquired at higher rates. Furthermore, the overall learning rates were not dependent on the lag (three versus seven trials)

between the exposure to a target word and the participant's production of the target word. Substantial knowledge was retained over a period of 20 minutes and six months.

In conclusion, the outcomes of this study provide insight in the learning rates of new, concrete L2 words when learners are unaware of taking part in a language learning study. Among other things, this line of research could be used to further identify those aspects of L2 learning that are relatively hard or easy to learn for untutored, immersed learners. In response, language courses for immigrants may shift their focus to those aspects of L2 learning for which tuition is indispensable. Other open questions concern the role of characteristics of the learner (e.g. age or proficiency), of the conversational partner (e.g. accent), or of the learning context (e.g. instruction). The new methodology that we presented will allow future researchers to investigate a large range of such questions on naturalistic, interactive L2 word learning in a highly-controlled immersion setting outside of the classroom.

Note

1. As per request of one reviewer, we included the participants' phonological working memory scores in our statistical models. However, these models soon failed to converge when we expanded the random-effects structure. A simple model that did converge showed that phonological working memory had virtually no effect on learning rates. Therefore, we continued the original, correlational analysis for investigating individual differences.

Data availability statement

The data and analysis script that support the findings of this study are available at https://github.com/johannadevos/NaturalisticL2WordLearning and at http://hdl.handle.net/11633/di.dcc.DSC_2017.00027_498.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723. doi:10.1109/tac.1974.1100705

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 20(5), 1063–1087.

Balota, D. A., Duchek, J. M., & Paullin, R. (1989). Age-related differences in the impact of spacing, lag, and retention interval. *Psychology and Aging*, 4(1), 3–9.

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, 57(1), 35–56.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. arXiv preprint, arXiv:1506.04967.

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

Bisson, M.-J., Van Heuven, W. J., Conklin, K., & Tunney, R. J. (2014). The role of repeated exposure to multimodal input in incidental acquisition of foreign language vocabulary. *Language Learning*, 64(4), 855–877.

Bordag, D., Kirschenbaum, A., Tschirner, E., & Opitz, A. (2015). Incidental acquisition of new words during reading in L2: Inference of meaning and its integration in the L2 mental lexicon. *Bilingualism: Language and Cognition*, 18(3), 372–390.

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, 20(2), 136–163.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3), 354–380. doi:10.1037/0033-2909.132.3.354

Crawley, M. J. (2007). *The R book*. Chichester: John Wiley & Sons. doi:10.1002/9781118448908

Dahl, A., & Vulchanova, M. D. (2014). Naturalistic acquisition in an early language classroom. *Frontiers in Psychology*, 5, article 329. doi:10.3389/fpsyg.2014.00329

De Bree, E. (2007). *Dyslexia and phonology: A study of the phonological abilities of Dutch children at-risk of dyslexia* (Doctoral dissertation). Retrieved from https://dspace.library.uu.nl/handle/1874/21522

De Groot, A. M. B. (2006). Effects of stimulus characteristics and background music on foreign language vocabulary learning and forgetting. *Language Learning*, 56(3), 463–506. doi:10.1111/j.1467-9922.2006.00374.x

De Groot, A. M. B. (2011). *Language and cognition in bilinguals and multilinguals: An introduction*. East Sussex: Psychology Press. doi:10.4324/9780203841228

De Groot, A. M. B., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, 50(1), 1–56. doi:10.1111/0023-8333.00110

De Vos, J. F., Schriefers, H., & Lemhöfer, K. (2018). Noticing vocabulary holes aids incidental second language word learning: An experimental study. *Bilingualism: Language and*

Cognition. Advance online publication. doi:10.1017/S1366728918000019

De Vos, J. F., Schriefers, H., Nivard, M. G., & Lemhöfer, K. (2018). A meta-analysis and meta-regression of incidental second language word learning from spoken input. *Language Learning*, 68(4), 906–941. doi:10.1111/lang.12296

Doughty, C., & Williams, J. (1998). Pedagogical choices in focus on form. In C. Doughty, & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 197–262). Cambridge: Cambridge University Press.

Ellis, R. (1999). Factors in the incidental acquisition of second language vocabulary from oral input. In R. Ellis (Ed.), *Learning a second language through interaction* (pp. 35–61). Amsterdam: John Benjamins. doi:10.1075/sibil.17.06ell

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning*, 43(4), 559–617. doi:10.1111/j.1467-1770.1993.tb00627.x

Godfroid, A., Ahn, J., Choi, I., Ballard, L., Cui, Y., Johnston, S., … Yoon, H.-J. (2018). Incidental vocabulary learning in a natural reading context: An eye-tracking study. *Bilingualism: Language and Cognition*, 21(3), 563–584. doi:10.1017/S1366728917000219

Grey, S., Williams, J. N., & Rebuschat, P. (2015). Individual differences in incidental language learning: Phonological working memory, learning styles, and personality. *Learning and Individual Differences*, 38, 44–53. doi:10.1016/j.lindif.2015.01.019

Gullberg, M., Roberts, L., & Dimroth, C. (2012). What word-level knowledge can adult learners acquire after minimal exposure to a new language? *International Review of Applied Linguistics in Language Teaching*, 50(4), 239–276.

Huckin, T., & Coady, J. (1999). Incidental vocabulary acquisition in a second language: A review. *Studies in Second Language Acquisition*, 21(2), 181–193. doi:10.1017/s0272263199002028

Hulme, C., Maughan, S., & Brown, G. D. A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language*, 30(6), 685–701. doi:10.1016/0749-596x(91)90032-f

Hulstijn, J. H. (2003). Incidental and intentional learning. In C. Doughty, & M. Long (Eds.), *The handbook of second language acquisition* (pp. 349–381). Oxford: Blackwell Publishing.

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal*, 80(3), 327–339. doi:10.1111/j.1540-4781.1996.tb01614.x

Lapkin, S., Swain, M., & Smith, M. (2002). Reformulation and the learning of French pronominal verbs in a Canadian French immersion context. *The Modern Language Journal*, 86(4), 485–507. doi:10.1111/1540-4781.00157

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, 44(2), 325–343. doi:10.3758/s13428-011-0146-0

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10, 707–710.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, 4(1), 19–31.

Lotto, L., & De Groot, A. M. B. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning*, 48(1), 31–69. doi:10.1111/1467-9922.00032

MacLeod, M. D., & Macrae, C. N. (2001). Gone but not forgotten: The transient nature of retrieval-induced forgetting. *Psychological Science*, 12(2), 148–152. doi:10.1111/1467-9280.00325

Malt, B. C., & Sloman, S. A. (2003). Linguistic diversity and object naming by non-native speakers of English. *Bilingualism: Language and Cognition*, 6(1), 47–67. doi:10.1017/S1366728903001020

McGraw, I., Yoshimoto, B., & Seneff, S. (2009). Speech-enabled card games for incidental vocabulary acquisition in a foreign language. *Speech Communication*, 51(10), 1006–1023. doi:10.1016/j.specom.2009.04.011

Mondria, J.-A. (2003). An experimental comparison of the "meaning-inferred method" and the "meaning-given method". *Studies in Second Language Acquisition*, 25(4), 473–499.

Ortega, L. (2009). *Understanding second language acquisition* (1st ed.). London: Hodder Education. doi:10.4324/9780203777282

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, 2, 10. doi:10.3389/neuro.11.010.2008

Peterson, L. R., Wampler, R., Kirkpatrick, M., & Saltzman, D. (1963). Effect of spacing presentations on retention of a paired associate over short intervals. *Journal of Experimental Psychology*, 66(2), 206–209. doi:10.1037/h0046694

Powell, M. J. D. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *DAMTP 2009/NA06*. Retrieved from http://www.damtp.cam.ac.uk/user/na/NA_papers/NA2009_06.pdf

R Core Team. (2018). *R: A language and environment for statistical computing* [Computer software]. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Restrepo Ramos, F. D. (2015). Incidental vocabulary learning in second language acquisition: A literature review. *PROFILE Issues in Teachers' Professional Development*, 17(1), 157–166. doi:10.15446/profile.v17n1.43957

Robinson, P. (2002). Effects of individual differences in intelligence, aptitude and working memory on adult incidental SLA: A replication and extension of Reber, Walkenfeld and Hernstadt, 1991. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 211–266). Amsterdam: John Benjamins. doi:10.1075/lllt.2

Rosch, E. (1978). Principles of categorization. In E. Rosch, & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, 21(4), 589–619. doi:10.1017/s0272263199004039

Schmidt, R., & Frota, S. N. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 237–326). Rowley, MA: Newbury House.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363. doi:10.1177/1362168808089921

Shokouhi, H., & Maniati, M. (2009). Learners' incidental vocabulary acquisition: A case on narrative and expository texts. *English Language Teaching*, *2*(1), 13–23. doi: 10.5539/elt.v2n1p13

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*, *16*(3), 371–391. doi:10.1093/applin/16.3.371

United Nations, Department of Economic and Social Affairs. (2015). *Trends in international migrant stock: The 2015 revision*. Retrieved from http://www.un.org/en/development/desa/population/migration/data/estimates2/estimates15.shtml

Van Casteren, M., & Davis, M. H. (2007). Match: A program to assist in matching the conditions of factorial experiments. *Behavior Research Methods*, *39*(4), 973–978. doi:10.3758/bf03192992

Van Zeeland, H., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, *41*(3), 609–624. doi:10.1016/j.system.2013.07.012

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, *61*(1), 219–258. doi:10.1111/j.1467-9922.2010.00593.x

Williams, C. C., & Zacks, R. T. (2001). Is retrieval-induced forgetting an inhibitory process? *The American Journal of Psychology*, *114*(3), 329–354. doi:10.2307/1423685

Zinszer, B. D., Malt, B. C., Ameel, E., & Li, P. (2014). Native-likeness in second language lexical categorization reflects individual language history and linguistic community norms. *Frontiers in Psychology*, *5*, article 1203. doi:10.3389/fpsyg.2014.01203