Naturalistic Word Learning in a Second Language

Johanna F. de Vos

Het onderzoek in dit proefschrift is tot stand gekomen binnen het Vidi-project van Dr. K.M. Lemhöfer, toegekend door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (projectnummer 276-89-004).

ISBN
9789462841925
Schilderij cover
Frederik de Vos
Layout en coverontwerp
Dennis Hendriks / ProefschriftMaken
Drukken
ProefschriftMaken / Digiforce

© Johanna F. de Vos, 2019

Naturalistic Word Learning in a Second Language

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op woensdag 24 april 2019
om 12:30 uur precies

door
Johanna Françoise de Vos – Tamis
geboren op 3 april 1989
te Amsterdam

Promotor
Prof. dr. H.J. Schriefers

Copromotor
Dr. K.M. Lemhöfer

Manuscriptcommissie
Prof. dr. J.J.M. Schoonen
Prof. dr. A.M.B. de Groot (Universiteit van Amsterdam)
Prof. dr. J.H. Hulstijn (Universiteit van Amsterdam)

TABLE OF CONTENTS

1.
General introduction

The field of *second language acquisition* investigates how people acquire a second language: its words, grammatical rules and sounds, as well as other aspects, such as what is and what is not polite. With the term second language (L2) I mean any language that someone learns as a non-native language, whether it is in fact his/her first, second or third non-native language. Second language acquisition is studied from multiple perspectives, for example what factors influence how many words someone can learn from reading a text (e.g., Godfroid et al., 2018), or how age affects L2 learning (e.g., Granena & Long, 2013). Many of these studies have applied goals, like finding the most effective way to train grammar or vocabulary (e.g., Van den Broek, Takashima, Segers & Verhoeven, 2018), or to develop tests of language proficiency (e.g., Harsch & Hartig, 2016). The outcomes of such studies can be used, among other things, to facilitate the L2 acquisition process of learners in the L2 classroom.

While language acquisition often takes place as the result of formal instruction (for example in language classes), many people also learn an L2 simply by being exposed to it. L2 learners may live and/or work abroad in a country where the L2 is spoken, or they might speak the L2 on a coffee date with a friend. They could encounter the L2 as the language of instruction in the classroom, while studying a topic other than the L2 itself. For example, many universities in the Netherlands are offering degrees that are taught in English. Learners can also encounter and acquire the L2 when travelling. Such naturalistic language learning outside of a tutored context is the focus of this thesis. Specifically, this thesis concerns naturalistic L2 word learning.

Formally, naturalistic language learning has been defined with various terms, including "informal and unstructured" (Mitchell & Myles, 2004, p. 6), "meaning-driven" (De Graaff & Housen, 2009, p. 726), "non-instructed" (De Graaff & Housen, 2009, p. 728), "within the target language community" (Howard, 2005, p. 495), and with "no classroom contact" (Dewaele, 2005, p. 542). While all pointing in the same direction, these definitions show that what is considered the very essence of naturalistic learning varies between researchers. For some, naturalistic learning by definition happens outside of schools (e.g., Dewaele, 2005). For others, the communicative aspect is of central importance (e.g., Howard, 2005). What I consider essential for naturalistic learning is contact with the target language in the absence of language-focused tuition. This does not preclude the possibility that learners have received L2 instruction at some point in the past, or will in the future. After all, most learners acquire the L2 due to a combination of instruction and naturalistic exposure. However, in this thesis I have exclusively focused on L2 learning in the absence of tuition.

I wished to study this phenomenon in a variety of contexts. These contexts usually adhered to most of the above definitions, but not all at the same time. Still, all of the above definitions of naturalistic learning were covered in at least one of my studies. Before introducing my own research, I will present some examples of other studies on naturalistic L2 learning.

1.1 NATURALISTIC L2 LEARNING: A RESEARCH IMPRESSION

An early study comes from Snow and Hoefnagel-Höhle (1978), who followed 51 native English speakers in different age groups who had just moved to the Netherlands, and were simply "picking […] up" (p. 1115) the Dutch language at school or at work. The age groups were 3-5 years old, 6-7 years old, 8-10 years old, 12-15 years old, and adults. During a one-year period, the Dutch proficiency of the participants was tested three times, on nine different measures. These measures ranged from knowledge of Dutch sounds, grammar and vocabulary, to story comprehension and storytelling. Taking all age groups together, there was significant improvement on all measures, with only one exception (the measure of auditory discrimination did not improve between the second and third testing moment).

In other studies, the language development of students who studied abroad was monitored. Tanaka and Ellis (2003) found that the English grammar, listening skills and speaking skills of 166 Japanese students who spent 15 weeks abroad in the US improved significantly. Carroll (1967) found that American students who had studied abroad for one year (n = 411) had better L2 listening skills than students who had toured or had only spent a summer abroad (n = 696). In turn, the latter group outperformed students who had never been abroad (n = 977). Serrano, Tragant and Llanes (2012) followed 14 Spanish students in the UK during one academic year, tracking both their spoken and written English proficiency. For spoken language, they found that fluency and lexical richness already had improved after one semester, and that after the whole year, the students also made fewer lexical, morphological and syntactic errors. Only the syntactic complexity of their spoken language had not changed. Regarding written language, one semester was not enough to bring about any changes, but after one year the students had improved in all four aspects that were measured (fluency, syntactic complexity, lexical richness and accuracy).

There are more studies like the above, focusing on the effects of studying abroad on L2 language proficiency. However, perhaps they are not strictly naturalistic: We usually cannot distinguish between the improvement that is due to the L2 language classes that the students are taking in their host country, and the improvement that is due to every day, naturalistic exposure to the L2. This is also pointed out by Knoch, Rouhshad, Oon and Storch (2015), which is one of the few studies on the effects of studying abroad where the students did not (or hardly) get formal L2 instruction. As compared to other studies on the effects of studying abroad, the 31 participants in Knoch et al. seem to have booked less progress. They were students in Australia, mostly from Asian countries and all speaking English as an L2. At the beginning of their study abroad period, as well as three years later, they wrote an essay, twice about the same topic. Over this period of three years, only an improvement in fluency could be detected (measured as the total number of words produced), but not in accuracy, grammatical complexity and lexical complexity, and neither in global writing scores. I should point out that this study only concerned writing skills, so perhaps there was more improvement on other skills such as speaking and listening.

Going abroad is not the only way in which naturalistic language learning can potentially take place. Some learners are regularly exposed to an L2 in their own country, for example if they are taking classes that are taught in the L2, but whose content revolves around another topic. For instance, Aguilar and Muñoz (2014) followed 66 engineering students in Spain who were taught in English. Since the purpose of these classes was not so much learning English as it was learning about engineering, I consider such research into the use of English-medium instruction at universities to be relevant to the domain of naturalistic learning. After 60 hours of English-taught engineering instruction, the students had significantly improved in their English listening skills (but not grammar skills). In a cross-sectional study, Lei and Hu (2014) compared the English proficiency of Chinese Business administration students between those who studied for this degree in English (n = 64) and in Chinese (n = 72). A measure of their English proficiency at the start of the course was incorporated in the analysis in order to control for pre-existing group differences. No significant effect of studying in English versus Chinese on English proficiency could be detected over a period of one year.

1.2 NATURALISTIC LEARNING VERSUS EXPERIMENTAL CONTROL

As can be seen from this short review, naturalistic learning studies generally are observational. This has two disadvantages. The first is that it is difficult for researchers to control and document the L2 learning process of their participants: Researchers cannot control the naturalistic linguistic input the participants are exposed to, and often do not even know the exact characteristics of this input. It also typically is unknown to researchers whether the learning process was naturalistic throughout. Although Snow and Hoefnagel-Höhle (1978, p. 1115) state that there was "little or no formal instruction", their participants were working or going to school in the Netherlands for a whole year. It is likely that they received some form of language-related instruction during that period, whether they asked for such instruction or whether it was unsolicited. This is something that cannot be controlled by researchers. For the students in study abroad programmes, language classes may have been part of their curriculum. Ortega (2009, p. 6) indeed argues that second language acquisition often comes about as "a mixture of both naturalistic and instructed experiences."

The second disadvantage of observational research on naturalistic L2 learning is that a control group typically either is absent (as in most of the above-discussed studies), or that the assignment of learners to groups is non-random. For example, Carroll (1967) did not decide himself which students would go abroad and which students would stay in the US. While he found that the students who had studied abroad had the best L2 listening skills, it is possible that these students already had superior L2 skills before going abroad. At the very least, a pre-test is needed in such designs. Lei and Hu (2014) did a better job in controlling for pre-existing group differences in English proficiency by administering a pre-test and including its outcomes in their statistical model, but even so the possibility cannot be excluded that the students who chose to go abroad (e.g., Carroll, 1967), or chose to study in English (e.g., Lei and Hu, 2014) differed from the control group in other aspects, such as language learning

aptitude, intelligence or motivation. This is a common issue in naturalistic learning studies, and not easy to resolve, because big decisions such as whether or not someone should be 'assigned' to move to the Netherlands or to study abroad are not up to researchers.

In summary, while observational studies have the great advantage of naturalness (i.e., real-life linguistic input and interaction), they have the disadvantages that the characteristics of the input are not exactly known to researchers and not under their control, and that the assignment to experimental and control groups is non-random (if there even is a control group). The tension between naturalness and experimental control was also a challenge in this thesis. As will become clear in the below preview of the five studies in this thesis, we investigated naturalistic learning at different places on the 'naturalistic versus control' spectrum.

1.3 INCIDENTAL LEARNING AS AN APPROXIMATION OF NATURALISTIC LEARNING?

Content-wise, a general conclusion from the short review seems to be that naturalistic L2 learning indeed can take place (although it does not always). But because the naturalistic L2 studies are so diverse, and mostly conducted without a control group or random assignment to experimental conditions, it is difficult to draw more specific conclusions. Open questions regarding naturalistic L2 word learning are how much learning takes place on average, how the context for learning influences the learning outcomes, and which variables predict the amount of learning.

Answers to such questions can to some extent come from the domain of incidental learning, which conceptually shares a lot with naturalistic learning. The definitions, mechanisms and possible operationalisations of incidental learning will be detailed in Chapter 2, but as its name says, *incidental language* learning happens incidentally, in activities that are not primary aimed at language learning. In this first definition, the possible situations in which naturalistic language learning can take place, such as living abroad, being at university or speaking to a friend, are also all situations in which incidental word learning could take place. In a second, alternative definition, word learning is considered incidental if learners do not know that the activity they are engaged in will be followed by a vocabulary test (Hulstijn, 2003). This also applies to any naturalistic learning situation (in fact, usually there is no vocabulary test).

Only in its third sense, incidental learning cannot be equated with naturalistic learning. The difference is an important one. The third definition of incidental learning is "learning without intention" (Ortega, 2009, p. 94). This is the opposite of *intentional learning:* learning with an intention to learn, for example the intention to commit a word or grammatical rule to memory. Intentional learning will of course happen in language classes, but can happen in any situation when a learner decides that something is worth remembering, including when being on a coffee date or when travelling (i.e., also in naturalistic situations). In this sense, naturalistic learning will presumably always be a mix of intentional and incidental learning. However, the third definition of incidental learning is also considered as being the least

useful in research, because "intentions wax and wane and fluctuate" (Ortega, 2009, p. 94). Gass (1999, p. 320) also points out that we have no "direct access to what a learner is doing", and therefore we can never be sure whether a word was learned with or without intention.

Vice versa, naturalistic learning cannot always be equated with incidental learning. The five definitions of naturalistic learning cited at the beginning of this General introduction do not all apply to incidental learning. For example, there are many studies that focus on incidental learning in the L2 classroom, as long as it takes place in an activity that is not directly aimed at language learning, and/or as long as the learners do not know that they will be post-tested. On the other hand, one of the definitions of naturalistic learning is that it takes place outside of the L2 classroom.

In conclusion, while incidental and naturalistic learning cannot fully be equated, studies on incidental L2 word learning can at least inform us about the conceptual basis that these two types of L2 learning do share, namely learning that happens in situations that are not primarily aimed at language learning, and in which no vocabulary test is expected. This is relevant, because incidental learning has been studied much more extensively than naturalistic learning.

1.4 SUMMARISING THE CURRENT STATE OF KNOWLEDGE IN A META-ANALYSIS

Not only has incidental L2 word learning been more widely investigated in single studies, it has also been meta-analysed. A meta-analysis is a study in which the outcomes of earlier studies are combined in one overall analysis. Because different studies usually use different measures, meta-analyses work with standardised effect sizes. The usual outcome of a meta-analysis is the average effect size over all studies, typically marked as being small, medium or large (see Plonsky & Oswald, 2014). For example, Abraham (2008) found that L2 learners who had access to computer-mediated glosses while reading a text (i.e., who could look up word meanings on the computer) learned more words while reading than L2 learners who did not have access to glosses. The average effect size, over a total of 11 studies, was large. Huang, Willson and Eslami (2012) compared L2 incidental word learning between learners who completed an output task (such as a sentence writing or fill-in-the-blank task) to learners who only read a text, and found that output tasks were more beneficial for word learning. The average effect size of this contrast (over 12 studies) was large. Huang et al. also investigated the effect of five additional variables, including the type of output task. This was done by calculating the average effect size per output task type.

I initially wanted to open this thesis with a meta-analysis on naturalistic L2 word learning, in order to capture the current state of knowledge in one review study. However, upon starting this meta-analysis we found that the primary research (i.e., the individual studies that a meta-analysis would be based on) in the domain of naturalistic L2 word learning was scarce, extremely diverse, and often conducted with little experimental control. In other words, these studies did not seem suitable for our purpose. Therefore, we decided to run our meta-analysis over incidental learning studies in which incidental learning is defined according

to a characteristic that it shares with naturalistic learning, namely that language learning takes place in an activity that is meaning-focused. In this case, it was possible to draw from a pool of well-controlled, experimental studies, although the majority of them concerned the written domain (i.e., reading and writing). In this thesis, I specifically focused on incidental L2 word learning from spoken input, since language exchanges in daily life are most often spoken. By conducting a meta-analysis in this domain, we could create more insight in the effectiveness of learning from spoken input (which had not been meta-analysed before), across different learning contexts and learner populations.

We not only investigated the overall effectiveness of L2 word learning in such situations, but also zoomed in on five variables that may have influenced the magnitude of these effect sizes. To this end, we used the technique of meta-regression, which is a (multiple) linear regression analysis over effect sizes. One of the variables under investigation was the learning situation itself (e.g., listening only versus interaction). We also looked at the effect of the learners' age, and the type of vocabulary knowledge the learners were tested on (receptive versus productive knowledge). The final two variables had a methodological character. As pointed out above, I am concerned about the fact that no-input control groups are often lacking in naturalistic learning research, and wondered whether control group absence impacts effect size magnitudes in incidental learning studies. Therefore, we compared effect sizes between studies that did and did not include such a control group. In a similar line of reasoning, we compared effect sizes between studies that used pre-test to post-test gain scores to control for participants' pre-existing knowledge, and studies that did not. This focus is in line with current efforts to increase the methodological quality of L2 research (e.g., Norris, Ross & Schoonen, 2015; Plonsky, 2013).

1.5 NATURALISTIC LEARNING AND EXPERIMENTAL CONTROL?

As I mentioned, the existing literature on naturalistic learning mostly consists of observational research. This kind of research shows the result of the learning process, but it does not inform us of the details of L2 acquisition as it happens in the moment. In this respect, the experimental incidental learning literature complements the naturalistic learning research, because in experiments researchers can observe the learning process itself. We summarised the existing research on incidental L2 word learning from spoken input in a meta-analysis, which provides insight in overall L2 word learning rates and some variables that influence these. However, because the meta-analysis was based on work done by other researchers, it did not allow us to manipulate certain variables that we were interested in, and to zoom in on learning as it happens in the moment. In addition, as explained earlier, incidental learning and naturalistic learning are not exactly the same.

Therefore, we wished to study naturalistic L2 word learning in a controlled setting. We set out to develop an experimental approach in which the learning would be as naturalistic as possible, while still retaining control over the quality and quantity of the language input. Perhaps this kind of learning does not conform to all the different definitions of naturalistic

learning, but it seems impossible, and undesirable, to conduct an "informal and unstructured" experiment. Therefore, we could not satisfy this particular definition of naturalistic learning in our experiments. They did comply with all other definitions given in the beginning of this General introduction: The L2 word learning was meaning-driven, non-instructed, took place within the target language community and outside the classroom.

The first experimental study is described in Chapter 3. We created a task and cover story which kept the participants completely unaware of the fact that they participated in an experiment about L2 word learning. In fact, they did not know the study had anything to do with language at all. Rather, they thought the experiment was about judging the price of different objects. All participants were German native speakers who had learned Dutch as an L2 and were immersed in a Dutch language environment, but they did not know that we were recruiting German native speakers exclusively. Unbeknownst to these participants, during the experiment we checked whether they would learn to produce the Dutch names of the objects whose price they were judging.

In this setting, we investigated the number of exposures that are needed for learning new L2 words, whether the recall of new L2 words is influenced by the number of other words that participants have encountered in the meantime, and whether word learning rates are influenced by the word's cognate status (i.e., whether or not the L2 word is related to its L1 translation in terms of form). We also checked how many of the newly learned words the participants could still recall after 20 minutes and six months.

In addition to successfully keeping the participants unaware of the study's purpose, another important innovation of this study was that we pre-tested the participants' existing knowledge of the target words in a test that was again concealed as a price judgment task. With the outcomes of this pre-test, we created a list of to-be-learned words for each participant on an individual basis. This novel procedure was executed using newly developed experimental software. In this way, all participants learned the same number of new words, while we took the participants' pre-existing Dutch word knowledge into account.

1.6 THE ROLE OF SWAIN'S OUTPUT HYPOTHESIS IN NATURALISTIC L2 WORD LEARNING

Having developed and successfully used this paradigm for studying naturalistic L2 word learning in the lab, we turned our attention to the Output Hypothesis (Swain, 1995). This is an important theory in the field of second language acquisition, which states that speaking the L2 (i.e., producing L2 'output') is beneficial for learning the L2. One of the hypothesised benefits of producing L2 output is that learners notice which words they do not yet know, and as a result they may pay close attention when they later hear these words being spoken by someone else. This is called the *noticing function of output*, or in other words *noticing the hole* (in your own vocabulary).

In Chapter 4 we describe why the current empirical evidence for the noticing function of output is not yet satisfying. Furthermore, the effect of noticing the hole has never been

studied when it comes to naturalistic L2 word learning. Therefore, even if this hypothesised effect can (to some extent) be supported in classroom studies, it is unknown whether it also benefits L2 word learning in naturalistic contexts. Using the paradigm from Chapter 3, we made participants aware of holes in their vocabulary knowledge. This was done by asking them to name objects in Dutch (again, in the context of a price judgment task), but not (yet) providing the participants with the actual names. This led them to notice the holes in their vocabulary. In a second, price judgment task, they were exposed to the words they had just noticed missing.

At the end of the session, we checked whether the participants in this experimental, noticing-the-hole condition had learned to produce the objects' names. We compared their scores to those of a group of control participants who had not been prompted to name the objects at the beginning of the session. Still, from the interviews we held with all the participants after they had finished the experiment it became clear that about half of the control participants had also noticed holes in their vocabulary, despite the fact that they were not asked to produce the object names until the post-test arrived. This was of course a natural consequence of the fact that researchers can never control participants' thoughts and awareness. We therefore analysed these participants' data as a separate group and in fact obtained new insights that we had not originally planned for.

1.7 NATURALISTIC LANGUAGE LEARNING AT UNIVERSITY

During the period in which I worked on the meta-analysis and the two experimental studies, opinion articles kept appearing in the Dutch media about the advance of English as the language of instruction in Dutch higher education. Proponents of English-medium instruction often argued that it would benefit students' English skills (e.g., Maex, 2017; Van Oostendorp, 2017). The presence of international students would increase the quality of classroom discussions, and enhance students' mutual understanding (e.g., Lizzini, Martijn, Munk & De Regt, 2017; Maex, 2017; Van Oostendorp, 2017). The possibility to teach in English would also help to attract the best researchers from all over the world (e.g., Sommers, 2017).

On the other hand, opponents argued that lectures and classroom interactions are not of the same quality when they are held in L2 English as compared to L1 Dutch (e.g., Hermans, 2017; Huygen, 2017; Kleinjan, 2017). English-medium education would also hinder students' Dutch language development, even though most Dutch students go into the Dutch job market after graduation (e.g., De Groot, Jurgens, Rawie & Verbrugge, 2018; Huygen, 2017; Maex, 2017). Furthermore, English-medium education could create an extra barrier for students from underprivileged backgrounds (e.g., Sommer, 2017; Teuling, 2017). What struck me was that many of these arguments, especially the ones regarding the effect of the use of English on students' learning processes and language development, seemed to be subjective rather than based on empirical research.

Radboud University also takes part in the trend of offering more and more study programmes in English. In the academic year 2016-2017, at the height of the above debate,

psychology students at Radboud University could choose for the first time whether they wanted to study in Dutch or in English. Except for the language in which the lectures and work groups were taught, and in which the students read some of the materials, the two tracks were identical. The students in the two tracks learned the same content, answered the same exam questions (albeit in a different language), and were taught by the same lecturers. This seemed to offer the perfect opportunity to add empirical weight to the ongoing debate about the use of English in higher education. At the same time, because the students' primary aim was to learn about psychology rather than refine their (Dutch or English) language skills, it was also a context in which naturalistic language learning could take place.

By comparing the lexical development of Dutch and German students in the Dutch and English tracks, we could investigate language learning that was informal and unstructured (while still being meaning-driven and non-instructed). While some of the students' contact with the study language did take place in classrooms, language classes were not part of the curriculum. Therefore, the language learning presumably was naturalistic. The definition that naturalistic language learning should take place within the target language community was satisfied for the students in the Dutch track, but not in the English track. This complete change in context, from a two-hour experiment in a strictly controlled lab environment, to a longitudinal observation of students as they progressed through their first year of study, allowed us to investigate naturalistic L2 learning from a different angle.

The study concerning the dual-language Psychology programme at Radboud University departs from the earlier chapters in several respects. To begin with, we did not collect our own data, but worked with data that were supplied to us by the Psychology educational institute. As a result, we could not control which demographic information was available about the participants. Among other things, we had to guess the participants' native language based on their nationality, which added some uncertainty. The participants also were not randomly assigned to conditions: It was of course not possible to force some students to take a three-year psychology degree in Dutch, and others to take the degree in English. These differences made the last study more challenging than the earlier studies, in terms of the conclusions that we were able to draw from the data.

Nevertheless, in Chapter 5 we examine the potential effect of the study language on the development of students' lexical richness in that language. Lexical richness refers to someone's productive vocabulary skills: the proportion of content words they use, the sophistication of their vocabulary, and its diversity. This is explained in more detail in Chapter 5. In Chapter 6, we examine the potential effect of study language on students' *study success*, with which we mean their grades, obtained number of European Credits (ECs), and drop-out rates. It is my hope that the last two chapters in this thesis do not only further our understanding of naturalistic lexical development in the L1 and the L2, but also enrich the present debate about whether or not the advance of the use of English in higher education should continue at the current speed.

1.8 THESIS OUTLINE

Chapter 2 contains the meta-analysis and meta-regression in which we studied incidental L2 word learning from spoken input. We looked at the magnitude of the overall learning effect, as well as the influence of age, learning task, test type, use of pre-test to post-test gain scores, and use of a no-input control group. The data and script for analysis can be found at https://github.com/johannadevos/MetaAnalysis.

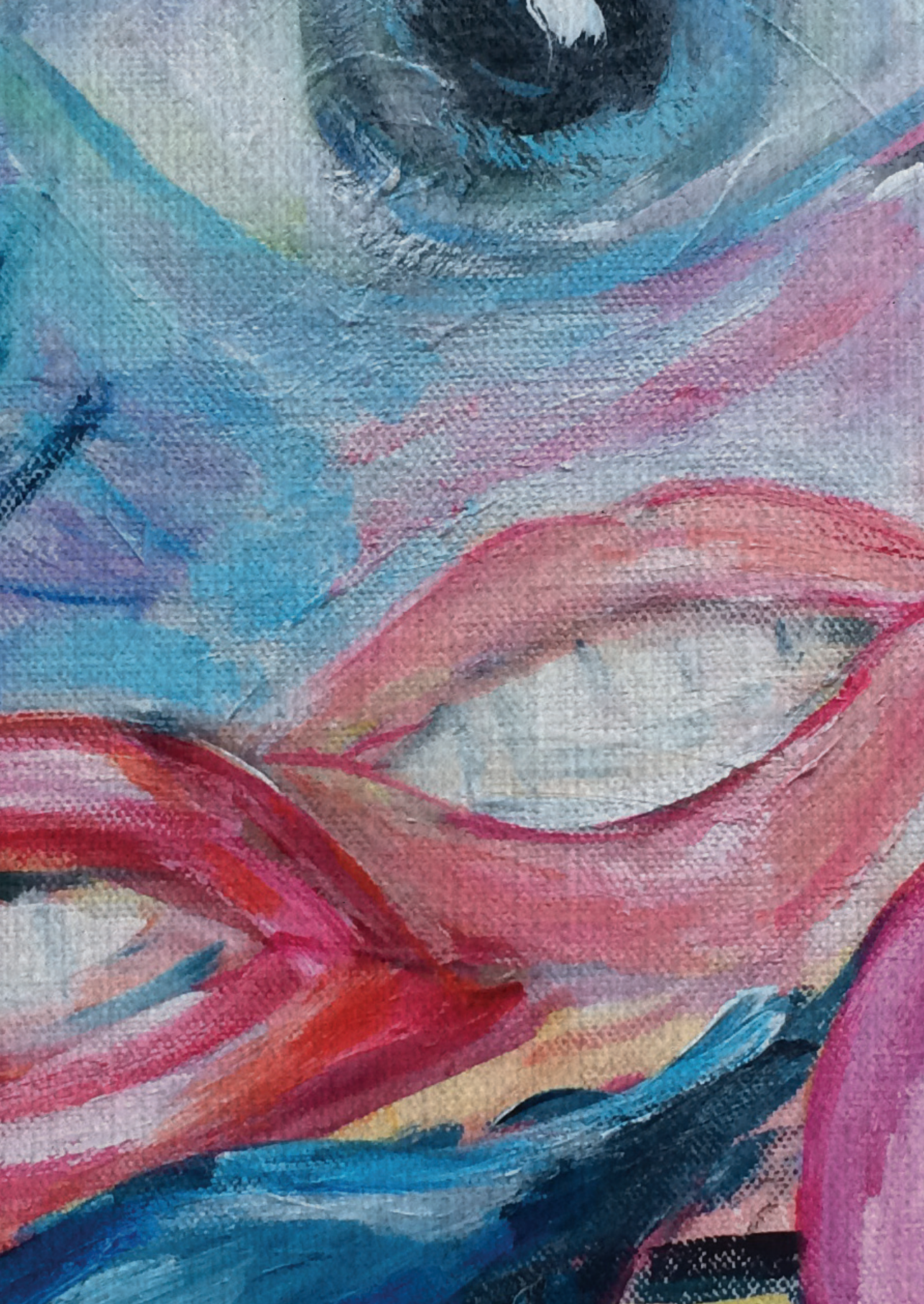
Chapter 3 describes the experiment we designed to study naturalistic L2 word learning in the lab. The participants were tested both during and after the learning task to see how many words they had picked up and remembered from the input. We also looked at the effects of cognate status, exposure frequency, the lag between exposure and production, and long-term retention. The data and script for analysis can be found at https://github.com/johannadevos/NaturalisticL2WordLearning.

In **Chapter 4**, we used an experimental set-up similar to that in Chapter 3, but this time we evaluated the hypothesised noticing function of output from Swain's (1995) Output Hypothesis. This was done by letting the participants in the experimental group notice holes in their vocabulary before providing them with the word forms they did not know. The data and script for analysis can be found at https://github.com/johannadevos/Noticing.

Chapter 5 describes the first part of the study we conducted with Dutch and German students in Nijmegen who studied psychology either in Dutch or in English. In this study, we looked at their lexical development during the first year of study. The data and script for analysis can be found at https://github.com/johannadevos/StudyLanguage.

The second part of this study can be found in **Chapter 6**. This time, we looked at the relationship between study language (Dutch or English), nationality (Dutch or German) and study success (i.e., grades, ECs and drop-out rates). The data and script for analysis can be found at https://github.com/johannadevos/StudyLanguage.

In **Chapter 7**, I summarise and discuss the outcomes of the studies presented in Chapters 2 to 6. Among other things, I will take a data-informed stance in the debate about the use of English in Dutch higher education.



2.

A meta-analysis and meta-regression of incidental second language word learning from spoken input

This chapter is based on:

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

ABSTRACT

We meta-analysed the effectiveness of incidental second language word learning from spoken input. Our sample contained 105 effect sizes from 32 primary studies employing meaning-focused word-learning activities with a total of 1,964 participants with typical cognitive functioning. The random-effects meta-analysis yielded a mean effect size of g = 1.05, reflecting generally large vocabulary gains from spoken input in meaning-focused activities. A meta-regression with three substantive and two methodological predictors also revealed that adult participants outperformed children in terms of word learning and that interactive learning tasks were more effective than non-interactive ones. Furthermore, learning scores were higher when measured with recognition than with recall tests. Methodologically, the use of a no-input control group seemed to protect against an overestimation of learning effects, as evidenced by smaller effect sizes. Finally, whether a pre-test to post-test design was used did not influence effect sizes.

2.1 INTRODUCTION

Second language (L2) learners living in a country where the L2 is used are often exposed to spoken L2 input in their daily lives. Even in situations that do not explicitly revolve around word learning, such incidental exposure can still result in the acquisition of new words. In the L2 classroom as well, words can be learned incidentally when learners listen to their teacher or peers without explicitly focusing on word learning. In short, L2 learners regularly find themselves in situations where incidental word learning from spoken input is possible.

Currently, however, little insight is available about the rate at which such learning takes place. Although research on incidental L2 word learning from spoken input exists (even if much less so than on learning from written input), in this prior work, researchers have typically compared the incidental spoken condition with another condition that is in some sense enhanced, for example, with additional written input (e.g., Brown, Waring & Donkaewbua, 2008) or with information about an upcoming vocabulary post-test (e.g., Montero Perez, Peters & Desmet, 2018). The spoken-only condition would then function only as a baseline and not be studied in itself. Therefore, it is practically unknown how effective incidental L2 word learning from spoken input is in an absolute sense (i.e., in comparison to a no-input condition). Experts' opinions on the topic also diverge. Ellis (1999) concluded that much of incidentally learned vocabulary comes from oral input, but Schmitt (2008) was more pessimistic and concluded that the literature on this topic mostly "points to a low uptake rate from listening exposure" (p. 349).

Our study had two goals aimed at increasing our understanding of the effectiveness of incidental L2 word learning from spoken input. First, we wished to systematically quantify this effectiveness by combining all available research in one meta-analysis. This allowed us to go beyond the conclusions that one can draw on the basis of individual studies. The resulting knowledge is relevant for teachers designing their curricula (e.g., should they include incidental learning activities with spoken input at all?), as well as for the many learners around the world who extensively rely on spoken input for their L2 acquisition.

Second, we aimed to investigate more closely how a selection of five variables affects incidental spoken L2 word learning. After all, it would be contentious to claim that there is one overall, all-embracing effect when many variables are known to influence L2 learning outcomes. At the same time, these variables have seldom been investigated within a single study of incidental L2 learning from spoken input. For example, of the 32 studies in our final sample, none compared children to adult language learners. To investigate the effect of age and other variables, we employed the technique of meta-regression, which enabled us to investigate variables whose levels varied between studies in one analysis. We could thus expand the existing knowledge about multiple variables using research that already existed.

2.1.1 Literature review

2.1.1.1 Defining incidental learning

Various definitions of incidental learning exist. It has often been defined by what it is not: Incidental learning would be "learning without intention, while doing something else" (Ortega, 2009, p. 94). A second definition specifically applies to incidental learning in the context of experimental research. According to this definition, incidental learning is dependent on the announcement of a post-test: When learners engage in an activity without the expectation of being tested afterwards, any resulting learning would be incidental (Hulstijn, 2003). A third definition is based on the nature of the activity that learners engage in: Learning can be considered incidental if it comes about as a "by-product" (Hulstijn, 2003, p. 362) of an activity that primarily revolves around meaning. This third definition was adopted for the current study. Incidental learning has often been contrasted with intentional learning, which is learning with intention, learning taking place in situations where learners know that they will be post-tested, or where the activities are explicitly focused on language learning.

2.1.1.2 Mechanisms of incidental learning

Although it has been well established that incidental learning is much less effective than intentional learning (e.g., Hulstijn, 2003; Schmitt, 2008), this does not necessarily mean that incidental learning is ineffective in itself, which is what we aimed to investigate in the current study. After all, it is incidental learning during non-tutored, everyday language use that turns learners into experienced L2 users. Multiple mechanisms have been proposed that can explain why and how incidental exposure to L2 words can result in learning.

In the first place, fast mapping might play a role. This notion, coined by Carey and Bartlett (1978), holds that children will generally try to map meaning onto new word forms that they encounter, using logical inference. They can construct this form–meaning link with as little as one exposure (making it fast) and even when no such link is explicitly provided in the input. Fast mapping is then driven by learners' innate curiosity for word learning. Adult language learners have been found to employ fast mapping as well, both when it comes to learning the meaning of non-words (e.g., Ramachandra, Rickenbach, Ruda, LeCureux & Pope, 2010) and incidental L2 word learning (e.g., Chapters 3 and 4).

Second, Hulstijn (2003), citing Eysenck (1982, p. 203), argues that the processing activities that learners engage in might influence learning rates more than their intentions. Building on the notions of depth of processing (Craik & Lockhart, 1972) and elaboration (Craik & Tulving, 1975), Laufer and Hulstijn (2001) developed the Involvement Load Hypothesis. Involvement is seen as consisting of three dimensions: need, search, and evaluation. Different L2 learning activities require different amounts of these motivational (need) and cognitive (search, evaluation) constructs, and activities that require a higher involvement from the learner are expected to lead to more learning. A meta-analysis by Huang, Willson, and Eslami (2012) found support for this hypothesis: Participants who completed an output task (which

supposedly was high in involvement) acquired more vocabulary than those who only read a text (which supposedly was low in involvement).

Third, learners can develop a curiosity or intention to learn words even when the activity that they engage in does not come with an announced post-test or is not explicitly focused on word learning. Thus, even in incidental learning contexts, learners can still decide to deliberately turn their attention to the input (Ortega, 2009), which can result in learning. It should be noted that this kind of learning would only be incidental according to the two definitions from Hulstijn (2003) but not the one from Ortega (2009).

2.1.1.3 Operationalising incidental learning

Because incidental learning is extremely difficult to operationalise when the learning-without-intention definition is used, we originally set out to find and analyse studies in which the learning was incidental according to the post-test announcement definition and the by-product definition. However, it turned out that using post-test announcement as a criterion was problematic too.

To begin with, for some studies in our sample it was unclear whether the post-test was announced (we contacted all authors to ask this, but not everyone replied). In addition, even in studies of which we knew that the post-test had not been announced, it could still have been expected by the participants. This already applied to all studies that used a pre-test: When learners are tested on unknown vocabulary and are exposed to this vocabulary afterwards, they probably expect a post-test. In addition, some studies used cycles of learning treatments and post-tests. For example, in Winke, Gass, and Sydorenko (2010), the participants watched a video twice and completed two vocabulary tests afterwards. This cycle was repeated three times. It is likely that after the first or second time, the participants knew that the vocabulary post-tests were coming.

For these reasons, we based the selection of studies on Hulstijn's (2003) by-product definition only and included studies in which the word learning treatment was presented as a meaning-focused activity, such as listening to an audiobook, watching a video, or performing an interactive task with a peer. In the context of the current study, we therefore speak of meaning-focused rather than incidental learning. For all included studies, we indicate in Table 2 (whenever possible) whether the post-test was announced. However, because there was much uncertainty with regard to whether the participants expected a post-test (even when it was not announced), we did not include this design feature as a variable in our analyses.

2.1.1.4 Meta-analysis and meta-regression in L2 research

For this study, we used the techniques of meta-analysis and meta-regression to analyse the learning outcomes of 32 studies. In general, meta-analysis allows researchers to calculate the weighted average outcome of a selection of studies. In the case of L2 word learning, such studies typically employ different tests, for example, a 10-item test requiring participants to

translate words from their L2 to their first language (L1) or an 18-item L1–L2 recognition test with four answer options per item. This means that such studies cannot be directly compared in terms of the average number of words that the participants learned. To compare them, the learning outcomes across studies need to be standardised by dividing the participants' gains by the standard deviation of their scores. This has been done in virtually all meta-analyses focusing on word learning (e.g., Abraham, 2008; Mackey & Goo, 2007; Montero Perez, Van den Noortgate & Desmet, 2013). By computing these average standardised learning effects over a multitude of studies, we were thus able to address the uncertainty regarding the effectiveness of incidental L2 spoken word learning.

Furthermore, we used meta-regression to investigate how five variables affect L2 incidental spoken word learning. Like 'ordinary' multiple regression, meta-regression is used to study how well the individual independent variables predict the dependent variable. The only difference is that, as in meta-analysis, the dependent variable is not the measurement originally used in the primary studies but a standardised effect size. Although meta-regression models technically are no different from other regression models, their use is still relatively rare in L2 acquisition research (for examples, however, see Goldschneider & DeKeyser, 2001; Li, 2010).

Instead, researchers often study predictor variables by splitting their data set by the levels of these predictor(s) and calculating separate effect sizes for all these subsets (e.g., Boulton & Cobb, 2017; Mackey & Goo, 2007; Montero Perez et al., 2013). Significance can be determined by considering whether the confidence intervals of the effect sizes of the subsets overlap (Mackey & Goo, 2007) or through *Q*-tests (Montero Perez et al., 2013). This has some disadvantages. In the first approach, no precise estimation of the significance level is obtained, and in both approaches, one needs to run a separate test for each contrast under investigation, which increases the chance of Type-I errors if no correction is applied. Li (2010) did use meta-regression, but with software that allowed "only one independent variable to be included" (p. 350). Nowadays, better software is available. Boulton and Cobb (2017, p. 382) argue that they did not use meta-regression because that would "mainly [be] suited to continuous [predictor variables]." However, categorical predictors can easily be included in a regression model through dummy coding or other forms of contrast coding. After all, an analysis of variance model is also a regression model with categorical predictors (Field, 2009).

In the present study, five predictor variables were analysed. Three were substantive: age of the participants, treatment, and mode of testing. In light of recent efforts to improve the quality of L2 research (e.g., Plonsky, 2013), two additional predictors focused on a methodological feature and concerned study design: whether a true control group was used and whether pre-test to post-test gain scores were computed.

2.1.2 Selected predictors of meaning-focused L2 word learning from spoken input

2.1.2.1 Age

The first predictor was the participants' age. While popular opinion often ascribes to the viewpoint that "younger is better" in L2 learning (Singleton & Ryan, 2004, p. 61), at the same time there is evidence that older learners might enjoy some advantages too, especially in word learning. Singleton and Ryan summarised the evidence regarding word learning as follows: There seems to be an advantage for older over younger participants in both short-term and long-term instructional studies as well as in short-term naturalistic (e.g., immersion, immigration) studies. Younger participants, however, eventually tend to overtake older participants in long-term naturalistic studies. One explanation for these findings comes from Paradis (2004) who suggested that vocabulary learning is not susceptible to a critical period for language learning (which would favour younger learners) because it relies on declarative memory. Thus, older language learners might benefit from their cognitive maturity when it comes to word learning.

In the current study, we compared the ability of L2 learners of different ages for meaning-focused L2 word learning from spoken input. As mentioned earlier, none of the studies in our sample had investigated age. There have been other studies on age effects in L2 learning, such as Snow and Hoefnagel-Höhle (1978) and Granena and Long (2013), but these did not employ an intervention in which participants were incidentally exposed to L2 spoken input. Using meta-regression, we could investigate whether there was indeed an older-is-better effect in the intervention studies included in our meta-regression.

2.1.2.2 Treatment

The second predictor was the learning treatment or intervention. A wide variety of activities can support meaning-focused learning from spoken input, such as listening to stories or audiobooks, watching videos, or interactive tasks such as solving a puzzle together. However, comparisons between such treatment types are relatively rare, especially between task-based and non-task-based learning activities. As with age, the technique of meta-regression is relevant for analysing treatment effects because treatment type does not need to be manipulated within a single study. In addition, the outcomes of different studies focusing on different learning tasks have been inconclusive. For example, Ellis, Tanaka, and Yamazaki (1994) found an advantage for tasks that involved negotiation between a participant and his/her L2 conversational partner (as compared to no negotiation), but Ellis and He (1999) did not find such an advantage. In this case, meta-regression can also provide a solution because combining the outcomes of multiple studies in one analysis should increase the power to detect differences between different task types.

Specifically, we compared the effectiveness of four different treatment types, which we chose because they have all been investigated regularly in primary studies. We compared audio treatments in which the input was presented auditorily only, for example, through audiobooks, with audiovisual treatments where the target words were also visually supported,

for example, through pictures (e.g., Brown et al., 2008) or through video (e.g., Montero Perez et al., 2018). We also included two task-based treatments. These also contained audio and visual input, but in addition there was the element of a meaning-focused task. Within the task-based treatments, we made a distinction between the presence and absence of interaction (+/– interaction) between the participant and a conversational partner. This has been commonly manipulated in task-based research (e.g., De la Fuente, 2002; Ellis & He, 1999). Thus, each of the four treatments under investigation was different from the previous one in a single aspect. This is schematically illustrated in Table 1.

Table 1. A schematic representation of the characteristics of the four treatments investigated.

→ Characteristics ↓ Treatment	Audio input	Visual input	Task-based	Interaction
Audio	✓	✗	✗	✗
Audiovisual	✓	✓	✗	✗
task/-interaction	✓	✓	✓	✗
task/+interaction	✓	✓	✓	✓

Note. ✓ = characteristic present; ✗ = characteristic absent.

Because this meta-analysis and meta-regression concerned L2 word learning from spoken input only, we excluded treatments where the spoken input was accompanied by text, such as a written transcript, glosses, L1 subtitles or L2 captions. In the case of L2 captions, it would be unclear whether the participants learned from the spoken or written version of the input. Treatments with L1 subtitles or translations were excluded because they remove the need for participants to deduce the meaning of a new word from the context or a visual scene, which we considered an essential part of learning from spoken input. In addition, the reading process could have interfered with the listening process. For meta-analyses on the effects of subtitling, see Montero Perez et al. (2013), for meta-analytic work on glossing, see Abraham (2008), and Yun (2011).

2.1.2.3 Mode of testing

From both anecdotal and scientific evidence, it is known that when learners are asked to remember previously learned words, open questions (recall) are generally more challenging than multiple-choice questions (recognition) (e.g., Donkaewbua, 2009; Montero Perez, Peters, Clarebout & Desmet, 2014). While the first two predictors (age and treatment) presumably mainly influence the learning process itself, testing usually only takes place after the learning phase has been completed. Therefore, rather than influencing learning success, it reflects the depth at which the newly acquired word knowledge can be processed. Given the important role that testing instruments play in L2 research and in education contexts, we chose to include mode of testing (recall versus recognition) as our third predictor, investigating the question of whether effect-size magnitude depends on testing mode.

2.1.2.4 Methodological predictors: Gain scores and control group

Finally, we included two methodological predictors relating to study design. When designing any word learning study, one has to ensure that the vocabulary knowledge displayed in a post-test can rightfully be attributed to the treatment and not, for example, to pre-existing knowledge of the target words that the participants already possessed. One solution for acknowledging pre-existing knowledge is to calculate pre-test to post-test gain scores. While this has the advantage that pre-existing knowledge can be controlled for with great precision, there are also multiple disadvantages associated with this approach, especially in studies that target incidental learning.

First, the presence of a pre-test might lead participants to also expect a post-test, making it questionable whether any potential learning should be considered incidental (this is why we instead concentrated on meaning-focused learning in this study). According to Schmitt (2008, p. 341), intentional vocabulary learning "almost always leads to greater and faster gains" than incidental learning. Second, as pointed out by Bisson, Van Heuven, Conklin and Tunney (2014b), as well as by Nation and Webb (2011), a pre-test also highlights the target words, perhaps causing learners to pay more attention to these words in later input than they would otherwise. For these reasons, studies making use of pre-test to post-test designs might be expected to yield higher effect sizes than studies using non-words or an independent control group to control for pre-existing knowledge. The inclusion of a predictor in the meta-regression for the use of gain scores should shed more light on such potential unwanted effects in L2 word learning studies.

A different approach is the use of a true control group, that is, participants who are not in any way exposed to the target words but who take the same tests as the experimental participants. In this way, researchers can again control for (group-level) pre-existing knowledge, although in a less precise and individual manner than when using a pre-test. In addition, researchers can control for any learning that might happen just as the result of taking tests, spontaneous fluctuations in behaviour, the passing of time, and guessing. The latter is especially relevant when the L1 and the L2 are closely related (see Chapter 3). Thus, in studies without a true control group, effect sizes might be overestimated, and lower effect sizes might be found in studies with a true control group. A true control group predictor was included in the meta-regression to investigate this.

2.1.3 The current study

To summarise, this is the first meta-analysis and meta-regression to bring together all literature on meaning-focused L2 word learning from spoken input. We documented the full research process to achieve maximal transparency and reproducibility. The data and script for analysis are publicly available at https://github.com/johannadevos/MetaAnalysis. The technical details that could not be included in this chapter due to space limitations can be found in Appendices A and B at the end of this chapter. In addition, the published version of this chapter on the *Language Learning* website is accompanied by Online Appendices S1

and S2, which are Excel files with details of the included and excluded studies (S1), and with details of the effect size calculations (S2). The study addresses the following questions:

1. What is the overall effectiveness of meaning-focused exposure to spoken input in L2 word learning?
2. How strongly is this effectiveness influenced by participants' age, type of treatment, and mode of testing?
3. Are effect sizes dependent on such study design features as the use of gain scores and the use of a true control group?

2.2 METHODS

2.2.1 Search techniques and sources considered

Four electronic databases were comprehensively searched for relevant studies published until and including August 2017, with no lower limit set. Three of these were subject-specific databases: PsycInfo, Linguistics and Language Behavior Abstracts, and Education Resources Information Center. These databases extensively cover the fields of psychology, linguistics, and education, and they index research on L2 learning. In addition, we inspected the ProQuest Dissertations and Theses database, a collection of four million graduate dissertations and theses from around the world.

All databases were searched for sources whose titles contained at least one of the below search terms, in combination with *vocabulary* and/or *word**. The individual search terms are shown in below, separated by commas. *Acq** represents search terms related to acquisition, and *gam** refers to gaming. For instance, the first search term *incidental* was used in two searches: *incidental AND vocabulary* and *incidental AND word**. We also used search terms relating to written data (such as *subtitl**) because studies about these topics sometimes contained data that were relevant to our purposes (as explained below). This is the complete list of search terms:

incidental, natural*, implicit, listen*, spoken, oral, aural, task-based, interaction* AND learn*, interaction AND acq*, subtitl* AND learn*, subtitl* AND acq*, caption* AND learn*, caption* AND acq*, gam* AND learn*, gam* AND acq*.

In addition, we manually searched the reference lists of all included studies and of theoretical and review articles on incidental L2 word learning (Ellis, 1999; Gass, 1999; Huckin & Coady, 1999; Hulstijn, 2003; Restrepo Ramos, 2015; Schmitt, 2008) and inspected the online archives of the following journals (in September 2017): *Language Learning & Technology, System, Language Learning, Studies in Second Language Acquisition, and Computer Assisted Language Learning*.[1]

[1] All issues of *Language Learning & Technology* and *System* were inspected. We found no new studies that fit the inclusion criteria. Therefore, we only inspected the issues of *Language Learning*, *Studies in Second Language Acquisition*, and *Computer Assisted Language Learning* published after 2010. Again, this yielded no results that had not already been found in the database search.

We screened the titles and (in case of doubt) the abstracts of all search results in the above-described databases, reference lists, and online archives. If it seemed that at least one condition in a study met the below-defined inclusion criteria, we inspected the study's methods and results sections. We also included one of our own published studies (Chapter 4).

2.2.2 Inclusion criteria and search outcomes

The 10 inclusion criteria are listed below. Criteria 1 to 5 concerned the scope of a study, Criteria 6 and 7 ensured that a study was of acceptable scientific quality, and Criteria 8 to 10 ensured that all necessary data were available:

1. The target language was a second or foreign language to the participants.
2. The target vocabulary was not explicitly taught or studied, but embedded in a meaning-focused activity. The participants were not told in advance what the target vocabulary would be.
3. The participants had typical cognitive functioning.
4. The target word input (and, optionally, output) was exclusively spoken.
5. At least one dependent variable measured word knowledge.
6. It was clear to which intervention potential increases in word knowledge were attributable.
7. Pre-existing word knowledge was controlled for by the use of gain scores, a true control group or a very careful selection of target items with regard to the participants' pre-existing knowledge.
8. Standardised effect sizes could be calculated from the provided means and standard deviations or from raw data.
9. Information about the five predictors was available.
10. The full text was available.

The screening of titles and abstracts resulted in 319 sources (e.g., articles, monographs, dissertations) that seemed relevant. Thirty of these sources (9%) were found to meet all of the inclusion criteria and are listed under "Included studies" in Appendix S1 in the Online Supporting Information. The remaining 289 sources are listed under "Excluded studies," accompanied by the reason for their exclusion.

Oswald and Plonsky (2010) discuss the question of whether research that has not undergone peer review should be included in meta-analyses, and deem the use of both peer-reviewed and non-peer-reviewed work acceptable (pp. 91–92). Including only peer-reviewed studies has the advantage that all studies can be expected to be of an acceptable scientific quality (Burnham, 1990, cited in Oswald & Plonsky, 2010). The advantages of also including non-peer-reviewed studies include an increase in statistical power and more robust results (Oswald & Plonsky, 2010). Boulton and Cobb (2017) included non-peer-reviewed research in

their meta-analysis because they wanted to obtain a sample as comprehensive as possible, and because they considered the peer-review process to be "highly subjective" (p. 354).

Given the relative scarcity of studies that met our inclusion criteria (in combination with our number of predictors), we also chose to include non-peer-reviewed research. To guard the quality of the included studies, we implemented several methodological checks. In line with Criterion 6, post-test data were only considered if we could determine that any potential learning could be attributed to the treatment (and not to earlier post-tests). Using Criterion 7, we checked whether participants' pre-existing word knowledge could be accounted for.

2.2.3 Characteristics of the sample

The included 30 sources contained relevant data from 32 studies (of which 24 were peer-reviewed), with a total of 44 independent treatment groups that were of interest to us and eight true control groups. The total number of participants over all included groups was 1,964. The mean number of participants in the independent treatment groups was 36 (*SD* = 39, range = 8–187), and in the control groups it was 41 (*SD* = 25, range = 11–82). Ten studies were published in the 1990s, five in the 2000s, and 17 in the 2010s. The 32 primary studies are described in Table 2, with additional information provided in Appendix S1 in the Online Supporting Information. The participants' proficiency in the target language covered the full spectrum, ranging from no pre-existing knowledge to high proficiency. All of the studies employed custom-made vocabulary tests containing the target words that the participants had been incidentally exposed to during the intervention. Two of the studies used non-words as targets (see Table 2). Additional information at the effect-size level (e.g., sample size, treatment type, and mode of testing) is available in Appendix S2 in the Online Supporting Information. Because many of the studies in our sample did not report the participants' age, in Table 2 we report age groups instead (see 2.2.3.1 for more details).

Table 2. Basic information about the 32 included studies.

Study	Age group	Gain scores	Control group	Post-test announced?	L1	L2
Al-Homoud (2008): Study 2	University?	Yes	No	?	Arabic	English
Aldera & Mohsen (2013)	University	Yes	No	No	Arabic	English
Baltova (1999)	High school	Yes	No	No?	Mostly, English in combination with another language	French
Birulés-Muntané & Soto-Faraco (2016)	University	No	Yes	No?	Catalan, Spanish or Italian	English
Bisson et al. (2014a)	University	No	Yes	No	English	Dutch
Brown et al. (2008)	University	No	No	No	Japanese	English (but non-word targets)
De la Fuente (2002)	University	No	No	Yes	English	Spanish

De Vos et al. (2018) (i.e., Chapter 4)	University	No	No	No	German	Dutch
Donkaewbua (2009)	University	Yes	Yes	No	Thai?	English
Duquette (1993)	University	Yes	Yes	Yes	English	French
Ellis & He (1999)	University	No	No	No	Various, mostly Asian	English
Ellis & Heimbach (1997)	Kindergarten	No	No	No	Japanese, Tagalog, Thai	English
Ellis et al. (1994): Saitama school	High school	No	No	No	Japanese	English
Ellis et al. (1994): Tokyo school	High school	No	No	No	Japanese	English
Gullberg et al. (2012): Exp. 1	University	No	No	No	Dutch	Mandarin
Gullberg et al. (2012): Exp. 2	University	No	No	No	Dutch	Mandarin
Hatami (2017)	University	No	Yes	No	Farsi	English
Hsu et al. (2013)	Elementary school	Yes	No	Yes	Language(s) of Taiwan?	English
Karakaş & Sariço-ban (2012)	University	Yes	No	No	Turkish?	English
Koolstra & Beentjes (1999)	Elementary school	No	Yes	No	Dutch?	English
Medina (1990)	Elementary school	Yes	No	No	Spanish	English
Montero Perez et al. (2014)	University	Yes	No	No	Dutch	French
Montero Perez et al. (2018)	University	No	No	Yes and no (manipulation)	Dutch	French
Nagata et al. (1999)	University	No	No	No	Japanese	English
Rodgers (2013): Study 2	University	Yes	Yes	No	Japanese	English
Sydorenko (2010)	University	No	No	No?	English (all but one, who spoke Canton-ese)	Russian
Toya (1993)	University	Yes	No	Yes	Japanese	English
Van Zeeland & Schmitt (2013)	University	No	No	No	Various	English (but non-word targets)
Vidal (2011)	University	Yes	Yes	No?	Spanish?	English
Winke et al. (2010)	University	No	No	No?	English (all but one, who spoke Kannada)	Spanish
Yeung et al. (2016)	Kindergarten	Yes	No	No	Cantonese	English
Yuksel & Tanriverdi (2009)	University	Yes	No	No?	Turkish?	English

The primary studies in our sample for the most part did not aim at answering the same research questions as we did in this study. What represented the treatment condition of interest for us (i.e., meaning-focused exposure to spoken-only L2 input) was often used as a control condition for studying the effects of subtitles, captions or glosses on incidental L2 word learning in the primary studies. This explains the low number of true control groups in our sample: The primary studies often achieved statistical control through other comparisons. For the same reason, we did not create a funnel plot of the effect sizes. Funnel plots are common in meta-analyses to indicate whether publication bias might be present concerning a certain effect. However, because the large majority of the primary studies in our sample had not investigated the effectiveness of meaning-focused exposure to spoken L2 input as compared to a no-input condition (like we did in the current study), their publication status was not dependent on the effect size(s) that we extracted from these studies.

2.2.3.1 Age

The studies included participants in different age ranges. Because only a minority of studies reported the participants' mean age (making continuous regression impossible), we created age groups based on the type of education that participants were enrolled in. Two studies were conducted with children in kindergarten (yielding five effect sizes), three with elementary school students (10 effect sizes), three with high school students (16 effect sizes), and 24 with university students (74 effect sizes). Due to the low number of effect sizes, we grouped children in kindergarten and elementary school together for our analysis. The confound between age and education type will be addressed in the Discussion.

2.2.3.2 Treatment

We distinguished four types of activities in which participants in the treatment groups engaged:

1. Audio (eight studies, 28 effect sizes): Participants listened to storybook reading, academic lectures or audiobooks.
2. Audiovisual (18 studies, 32 effect sizes): Participants received auditory input, but also visual support in the form of pictures or video.
3. Task/–interaction (six studies, 25 effect sizes): Participants listened to a speaker (a physically present teacher or peer), had materials that provided visual support, and engaged in a meaning-based task with these materials.
4. Task/+interaction (six studies, 20 effect sizes): The same as the previous activity, but participants also interacted with the speaker (e.g., they could ask questions). This sometimes, but not always, involved prompted production of the target words.

Appendix S2 in the Online Supporting Information shows which treatments and modes of testing were used in which studies. This information is not included in Table 2 because

treatment and mode of testing sometimes varied within a study, whereas Table 2 includes study-level information only.

2.2.3.3 Mode of testing

Testing the newly acquired word knowledge was done either by assessing the recognition of words through multiple-choice questions (51 effect sizes) or their recall via open questions (54 effect sizes). Furthermore, responses could be required in the L1 or L2. Recall was always meaning-based (e.g., a translation test) and recognition usually so, although in a minority of cases it was form-based (e.g., lexical decision). Post-tests could be administered in spoken or written form. Some studies employed one measurement type only; others employed multiple types. Similarly, some studies used one post-test, and others tested participants after various periods of retention. Thus, we could calculate 105 effect sizes from only 44 treatment groups. The dependency among these effect sizes is discussed in the Analysis section (2.2.6).

However, in the case of repeated post-testing, we only used these repeated test results if the participants could not have learned from the earlier tests (see Nation & Webb, 2011). For example, Aldera and Mohsen (2013) first administered a multiple-choice vocabulary recognition test and then a meaning translation test. It could be the case that the participants' answers to the translation test were informed by the questions and answers that they had seen previously. This would be in conflict with Criterion 6, which states that it should be clear to which intervention potential increases in word knowledge are attributable. Thus, we discarded the outcomes of recall post-tests if these had been preceded by a recognition test using the same materials. Whenever this happened, it is noted in the Remarks column in the Effect sizes tab in Appendix S2.

2.2.3.4 Retention interval

Finally, it is commonly known that knowledge is gradually forgotten over time (e.g., Ebbinghaus, 1885/1913/2011), making the retention interval between exposure and testing an important variable in L2 research (e.g., Brown et al., 2008; Van Zeeland & Schmitt, 2013; Vidal, 2011). In our sample, the retention intervals ranged from immediately after the treatment (for about half of all effect sizes) to three months (Brown et al., 2008). However, it was not possible to use retention interval as a predictor because the results in the primary studies were often not reported as a function of the retention interval. For example, Rodgers (2013) had 13 teaching sessions (generally separated by one week, but sometimes two). Episodes of a television program were shown in Sessions 3 to 12 and the post-test took place in Session 13. This means that there were at least 11 weeks between the first episode and the post-test, but only one week between the last episode and the post-test. Thus, there was not one retention interval for all items, but the scores were not reported separately for each retention interval. Similar set-ups are found in Al-Homoud (2008), De la Fuente (2002), Medina (1990), and Yeung, Ng, and King (2016), and in Brown et al. (2008) and Vidal (2011) with regard to the delayed post-test.

2.2.4 Meta-analytic statistics

2.2.4.1 Calculating learning scores

Because we aimed to establish the effects of meaning-focused exposure to spoken input on L2 word learning, our dependent variable of interest was a learning score. We calculated learning scores using the data reported in the primary studies. A learning score always represented a contrast between test scores obtained with (or before) and without (or after) exposure to target words. Based on these study designs, we constructed four different learning score types:

1. Comparison within treatment group(s): Post-test scores to a fixed baseline
   The same level of pre-existing knowledge of the target items was assumed for all participants, and their post-test scores were compared to this assumed baseline. For example, Gullberg, Roberts and Dimroth (2012) exposed Dutch native speakers with no self-reported prior knowledge of Mandarin to a Mandarin weather forecast. Thus, any knowledge of Mandarin words demonstrated in the post-test should reflect learning as a direct result of the treatment. The post-test used in this study consisted of yes/no questions to probe word recognition. Of course, even with no existing knowledge of Mandarin, the chance of making a correct guess was 50%. Therefore, the baseline in this study was set at 50%.

2. Comparison within treatment group(s): Pre-test to post-test gain scores
   Learning scores were calculated as the gain scores between a pre-test and a post-test for one or more treatment groups, without using a control group. For example, Yuksel and Tanriverdi (2009) tested Turkish students' English vocabulary knowledge before and after they had watched an English movie clip.

3. Treatment group(s) compared to control group: Post-test scores only
   Learning scores were calculated by comparing the post-test scores of a treatment group (after exposure to target words) to the scores of a control group (without exposure to target words). For example, Koolstra and Beentjes (1999) tested the English vocabulary of Dutch children who had watched an English video and a comparable group of children who had not.

4. Treatment group(s) compared to control group: Pre-test to post-test gain scores
   This is a combination of the learning score Types 2 and 3. The pre-test to post-test gain scores of a treatment group were compared to the pre-test to post-test gain scores of a control group. For example, Spanish students in the treatment group in Vidal (2011) watched a videotaped academic lecture in between taking a pre-test and post-test whereas participants in the control group took only the pre-test and post-test.

2.2.4.2 Calculating effect sizes

We used Hedges' *g* as our effect-size measure. It was calculated by multiplying Cohen's *d* by Hedges' correction factor *J* (Borenstein, 2009), which accounts for the biasing effect of small sample sizes on Cohen's *d* (Hedges & Olkin, 1985). Cohen's *d* is a standardised effect-size measure and was calculated as one of the learning score types described above, divided by the standard deviation. The calculation of all measures is described in detail in Appendix A at the end of this chapter. This includes a description of the transformations that are required when combining data from studies with different designs (see Morris & DeShon, 2002). We applied such transformations to the data to ensure that the meaning of *g* was unaffected by the type of learning score it was based on. Appendix S2 in the Online Supporting Information contains all 105 effect sizes and their associated characteristics (age group, treatment, etc.).

Nevertheless, one might wonder whether effect sizes that are calculated with the data of a treatment group only (learning score Types 1 and 2) are of a different magnitude than effect sizes that are calculated by comparing a treatment and control group (learning score Types 3 and 4). Similarly, it is conceivable that effect sizes that are calculated by comparing pre-test to post-test gain scores (learning score Types 2 and 4) would differ from effect sizes calculated based on post-test scores only (learning score Types 1 and 3). To this end, the variables of control group inclusion (yes/no) and the use of gain scores (yes/no) were included in the analysis.

We also calculated the variance *v* associated with each effect size (see Appendix A for the statistical details). This variance characterises the distribution from which an effect size was sampled, and therefore is different from the variance characterising the distribution of the participants' scores that were used for computing the effect size. The effect-size variance can be used for various purposes (Morris, 2008), including to weigh effect sizes according to their inverse variance weight (Hedges & Olkin, 1985). In other words, effect sizes with larger sampling variances are weighted less. We followed this practice, which is recommended because it allows studies with more precise effect-size estimates to "contribute more to the meta-analytic average" (Oswald & Plonsky, 2010, pp. 95–96; also see Borenstein, 2009). The effect-size variance was also corrected for small sample bias by multiplying it with the squared correction factor *J* (Borenstein, 2009).

2.2.5 Interrater reliability

The information presented in Appendices S1 and S2 in the Online Supporting Information was extracted from the primary studies by two raters: the author of this thesis, and three graduate students in linguistics or psychology, who divided the work. For the quantitative data, such as means and standard deviations, the interrater agreement was 90%. For the qualitative data, including the levels of the five predictor variables, the interrater agreement was 96%. Following Boulton and Cobb (2017), the interrater agreement had been calculated by considering the number of discrepancies relative to the total number of cells. The first and

second raters resolved all discrepancies together through discussion and by rereading the primary studies.

2.2.6 Analysis

Research question 1 focused on the overall effectiveness of meaning-focused exposure to spoken input for L2 word learning. To address this issue, we explored several random-effects meta-analytic models with the *metafor* package (version 1.9-9; Viechtbauer, 2010) in R (version 3.3.1; R Core Team, 2018). While the 32 studies in our sample all met the inclusion criteria, they still represented a wide variety of study designs, L1–L2 combinations, materials, and so on. Therefore, it was expected that there would be heterogeneity among the effects that were estimated in each study. In other words, it was likely that the variation in effect sizes would exceed the variation that would have been expected due to random variables alone, such as participant sampling.

Such expected between-study heterogeneity can be statistically accounted for by including random intercepts at the study level. In our case, this means that the meta-analytic model estimated a unique effect for each of the 32 studies, rather than assuming that the 32 studies all estimated the exact same effect. In a similar vein, the true effects could be imagined to vary across the 105 effect-size samples (even if some of them came from the same study), for instance, as a function of the exact testing instrument (e.g., was it administered in spoken or written form, was it meaning-based or form-based, etc.). To accommodate this, random intercepts were also introduced at the sample level (see Konstantopoulos, 2011; Viechtbauer, 2017).

We investigated whether the inclusion of random effects indeed significantly improved model fit by comparing the models using likelihood ratio tests. To this end, we first ran a null model with neither fixed nor random effects. The null model was compared to a model with only random intercepts at the study level. This latter model was then compared to a model with random intercepts at both the study and the sample levels. The best-fitting of these models, hereafter referred to as Model 1, was used to answer Research question 1.

Research questions 2 and 3 asked how effect-size estimates for meaning-focused L2 spoken word learning are influenced by five predictors: (a) Age group: kindergarten/elementary school versus high school versus university, (b) Treatment type: audio versus audiovisual versus task/–interaction versus task/+interaction, (c) Mode of testing: recognition test versus recall test, (d) Use of gain scores: yes versus no, and (e) Use of a true control group: yes versus no. These five predictors were added as fixed effects to Model 1 (a random-effects meta-analytic model), yielding Model 2 (a mixed-effects meta-regression model). Parameters for both Model 1 and Model 2 were estimated with the restricted maximum likelihood procedure, which takes into account the number of fixed-effects parameters that are estimated.

Because Model 2 was a regression model, we investigated whether there were any effect sizes that exerted a disproportionate influence on the estimation of the model parameters

using Cook's distance (D_i, Cook, 1977). An early rule of thumb was that a Cook's distance larger than 1 should be considered reason for concern (Cook & Weisberg, 1982, cited in Field, 2009). A simulation by McDonald (2002) has shown that 0.85 may be a more appropriate guideline. Therefore, three effect sizes with $D_i > 0.85$ were excluded from the data set, and Model 2 was rerun on this reduced data set.

Finally, an important assumption in standard meta-analytic models is that the effect-size estimates are independent (Hedges, Tipton & Johnson, 2010). Because we calculated 105 effect sizes from 44 treatment and eight control groups, the sampling errors of the effect sizes were not always independent. The traditional approach in case of dependency is to compute the weighted average of all the effect sizes coming from the same treatment group (Borenstein, Hedges, Higgins & Rothstein, 2009). This solves the issue of dependency, but at the same time information is lost. This would be especially problematic in the meta-regression, for example, if a study used both recognition and recall tests. By averaging over these test types, the resulting effect size would represent a combined recognition/recall score, and therefore would be unsuitable for investigating the effect of Mode of testing. An alternative recommended by Hedges et al. (2010) for dealing with dependency is to explicitly model the correlations among the effect-size estimates coming from the same treatment group.

However, as also pointed out by Hedges et al. (2010), the correlations between the effect-size estimates needed to implement this strategy are often not available. In these cases, it is accepted that correlations from comparable studies be used to estimate the missing values (Borenstein, 2009). This is what we did in the current study. Reassuringly, Ishak, Platt, Joseph and Hanley (2008) found through simulation studies that "the results of multivariate meta-regressions were relatively insensitive to incorrect values of within-study correlations" (cited in Hedges et al., 2010, p. 45). We will also report a sensitivity analysis in which we investigated how robust the outcomes of our meta-analysis and meta-regression were when different correlational values were assumed (the methodological details are explained in Appendix B at the end of this chapter).

To take the correlation between the sampling errors of the effect sizes into account, we constructed the whole variance–covariance matrix of the sampling errors (see Appendix A for information on the variance calculations). The covariances were calculated according to the standard definition of covariance: $\text{covariance}_{a,b} = \text{correlation}_{a,b} \times SD_a \times SD_b$. The full covariance matrix can be found in Appendix S2 in the Online Supporting Information (tab: Covariance matrix); the formulas in the cells show which correlation was used or assumed. Alpha was set to .05.

2.3 RESULTS

2.3.1 Research question 1

Model comparisons showed that the model with random intercepts at the study level fit the data significantly better than the null model, $\chi^2(1) = 461.74$, $p < .001$ (Akaike's information criterion (AIC) dropped from 1214.60 to 754.87; lower scores indicate a better model). In turn,

the model with random intercepts both at the study and sample level (i.e., the effect-size level) fit the data significantly better than the model with random intercepts at only the study level, $\chi^2(1) = 536.03$, $p < .001$ (AIC dropped from 754.87 to 220.84). This means that the true effect sizes were both heterogeneous between and within studies, as we had expected. This best-fitting model was used for further analysis and is hereafter called Model 1.

Model 1 yielded a weighted average effect-size estimate of $g = 1.05$, 95% confidence interval (CI) [0.81, 1.28], $SE = 0.12$, $z = 8.77$, $p < .001$, based on the effect sizes of learning gains obtained in the primary studies. This learning gain was significantly larger than 0. Thus, L2 learners experience a significant increase in their vocabulary knowledge after meaning-focused exposure to spoken L2 input. The variance at the study level was estimated by the model as 0.31, while the variance at the sample level was estimated as 0.21. Profile likelihood plots of these variance components, included in Appendix B (Figure A), showed that we could be confident in these variance estimates. The intraclass correlation was 0.31/(0.31 + 0.21) = 0.59. This represents a fair correlation (Cicchetti, 1994) between effect sizes coming from the same study, which provides further justification for the inclusion of random intercepts at the study level.

2.3.2 Research questions 2 and 3

For reasons of space, the outcomes for Model 2 as computed on the full data set are given in Table A in Appendix B. In Table 3, we present the model outcomes after three cases with Cook's distance values of 2.52, 2.47 and 1.04 were excluded from the data set (see Analysis).[2] The beta estimates show the estimated increase or decrease in effect sizes (in standard deviation units) compared to the predictor level that was represented by the intercept.

For Age, no significant difference could be detected between participants in kindergarten/elementary school (the level represented by the intercept) and high school. However, there was a significant 0.92 increase in effect-size magnitude for participants in university compared to kindergarten/elementary school. The estimated effect-size difference between participants in university and those in high school was 0.92 – 0.74 = 0.18. We changed the order of variable levels in the model (this is called *relevelling*) to obtain test statistics for this contrast, which showed that the difference was non-significant, $\beta = 0.18$, 95% CI [–0.78, 1.13], $SE = 0.49$, $z = 0.36$, $p = .72$.

Treatment had four levels. As we did with the age variable, each of these treatment levels in turn was made the intercept. One contrast was significant, namely task/–interaction versus task/+interaction, $\beta = 0.63$, 95% CI [0.26, 1.00], $SE = 0.19$, $z = 3.37$, $p < .001$. Thus, participants learned more words when the learning task that they engaged in involved interaction with a conversational partner than when it did not. All other contrasts were non-significant (all $ps > .11$).

[2] These were both effect sizes from Yeung et al. (2016) and the recall effect size from Brown et al. (2008), respectively.

Table 3. Results from Model 2 after the exclusion of three influential cases.

Fixed effects	β	*SE*	*z*	*p*	Lower bound	Upper bound
Intercept	0.00	0.49	-0.01	.99	-0.96	0.95
Age: high school	0.74	0.59	1.27	.21	-0.41	1.89
Age: university	0.92	0.41	2.26	**.02**	0.12	1.72
Treatment: audiovisual	0.07	0.16	0.43	.67	-0.25	0.39
Treatment: task/-interaction	0.10	0.44	0.22	.83	-0.76	0.95
Treatment: task/+interaction	0.73	0.45	1.61	.11	-0.16	1.61
Testing: recognition	0.42	0.09	4.42	**< .001**	0.23	0.60
Gain scores: yes	0.03	0.32	0.11	.91	-0.59	0.66
Control group: yes	-0.47	0.23	-2.03	**.04**	-0.93	-0.02
Random effects	**Variance**	***SD***				
Intercept (study)	0.49	0.70				
Intercept (sample)	0.05	0.23				

Note. k = 102.[3] The intercept represents the following combination of variable levels: Age = kindergarten/elementary school, Treatment = audio, Testing = recall test, Gain scores = no, Control group = no). Significant *p*-values are printed in bold.

Using a recognition test significantly increased effect sizes with 0.42 standard deviation units relative to using a recall test. No difference was found between studies that controlled for previous knowledge through gain scores and studies that did not use gain scores. On the other hand, studies that used a control group receiving no target word input yielded significantly lower effect sizes than studies that did not use a control group, with a difference of 0.47 standard deviation units.

Table 3 also shows that variance at the study level was estimated as 0.49, while variance at the sample level was estimated as 0.05. Profile likelihood plots again showed that we could be confident in these parameter estimates (see Figure B in Appendix B). For Model 2, the intraclass correlation was estimated as 0.49/(0.49 + 0.05) = 0.91. This represent a very high correlation between effect sizes coming from the same study. In other words, almost all of the variance was between studies, not within.

An inspection of Table A in Appendix B revealed that the outcomes of Model 2 carried out using the reduced data set (with three influential cases excluded) and unreduced data sets were very similar. The more salient differences are that the age effect was slightly more pronounced in the reduced data set and that the control group effect was non-significant in the unreduced data set. However, the *p*-value for the latter effect only increased from .04 to .06, so this difference was not very substantial.

[3] k represents the number of effect sizes included in the analysis (in other words, the sample size).

2.3.3 Sensitivity analysis

For the studies that did not report the correlation coefficient(s) needed for our analyses, we had borrowed a correlation coefficient from the most similar study that we could find in our sample. To investigate how sensitive our outcomes were to variation in the magnitude of these correlations, we carried out a sensitivity analysis, which is reported in Appendix B. We observed only very small changes in the magnitude of the estimated predictor effects and their associated *p*-values, both for Model 1 and Model 2. The direction of all effects was preserved. This indicates that our data were robust to variations in correlation.

2.4 DISCUSSION

2.4.1 Meta-analysis

This study set out to quantify the overall effectiveness of meaning-focused exposure to spoken input for L2 word learning. We found an estimated average Hedges' *g* of 1.05. This means that participants on average improved their vocabulary knowledge by 1.05 standard deviations after meaning-focused exposure to spoken L2 input. Because the effect size is expressed in terms of standard deviation units, it is not possible to state in an absolute sense the number of learned words to which such an effect size would correspond.

Plonsky and Oswald (2014) present guidelines for the interpretation of standardised effect sizes in the context of L2 research. For between-group designs, they regard $d = 0.4$ as small, $d = 0.7$ as medium, and $d = 1.0$ as large (and g is equal to d except that it is corrected for small-sample bias). The studies included in our meta-analysis used both between-group and within-group designs. However, the above guidelines still apply to our results because we had already taken intragroup correlations into account when calculating effect sizes for within-group designs. Thus, the estimated effect size of $g = 1.05$ can be considered a large effect.

Finding this large effect could be considered surprising in light of the general pessimism that surrounds the effectiveness of learning from listening (e.g., Schmitt, 2008). However, while the linguistic input in all of the studies in our sample was exclusively spoken, three-quarters of these studies provided additional support for learning in the form of pictures, video or learning tasks. Another explanation for the large effect size might be that the participants in some studies were aware of the upcoming post-test and therefore perhaps paid more attention to the target words. This may have resulted in larger effect sizes than would be found in studies in which the learning was incidental also according to the post-test announcement definition.

To put our finding in perspective, the outcome of the one other meta-analysis (that we know of) in which absolute L2 word learning gains were studied should be considered. Mackey and Goo (2007) studied improvements in vocabulary and grammar before and after an interactive treatment. Averaging over these two domains, they found an effect size of $d = 1.09$. This is very close to our estimated effect size of $g = 1.05$. It is currently impossible to say how absolute learning from spoken input compares, for example, to learning from written

input, because we believe no such meta-analysis has been conducted yet—this would be an important avenue for future research.

In conclusion, meaning-focused treatments for L2 word learning may be more effective than has previously been thought. Still, it is possible that the magnitude of our overall effect size was mostly attributable to specific subsections in our data set, such as the studies using task-based learning. The meta-regression provides more insight into this.

2.4.2 Meta-regression

We investigated whether the effectiveness of meaning-focused exposure to spoken L2 words is predicted by three substantive and two methodological variables.

2.4.2.1 Age

A positive Age effect was found: Participants in university significantly outperformed children in kindergarten and elementary school with a medium-to-large effect size. Effect sizes for high school participants also were higher than for younger children, but this difference did not reach significance (potentially due to small sample sizes for both age groups). Multiple explanations for the superiority of university students over kindergarten and elementary school children are conceivable. To begin with, older learners have more experience in language learning. For example, they might know more strategies to derive word meaning from context. In addition, they possess a higher degree of cognitive control (Craik & Bialystok, 2006), making it easier for them to focus on a task. There are also potential explanations based on confounds between age group and other variables.

First, the older groups on average had more years of experience with the L2, and, relatedly, typically seemed to have been more proficient (see Appendix S1 in the Online Supporting Information). In turn, learners with higher proficiency levels are known to learn vocabulary faster (Vidal, 2011). If more primary studies with adult learners are conducted in which the participants have no prior experience with the target L2 (such as Gullberg et al., 2012), a future meta-regression could circumvent this issue. In any case, age, proficiency levels, and years of experience with the target language should be clearly reported in all primary studies (this was not the case in our sample) so that future meta-analysts can better control for these variables.

Second, the adult participants perhaps also had a higher motivation to learn words. With regard to the classroom studies, the adults had chosen to enrol in language classes. For the children, language study simply was part of the school's curriculum. With regard to the laboratory studies, presumably the adults had volunteered to participate whereas the children would have been signed up by their school or parents. If more primary studies are conducted with adult participants other than self-selected language learners or if a study's language learning aspect is hidden from the participants (such as in Chapters 3 and 4), this confound could be alleviated.

Third, different average intelligence quotient levels can be expected between the general school population and university students. Because "foreign language aptitude partially overlaps with traditional intelligence" (Ortega, 2009, p. 165), this may also have influenced the results, implying an urgent need for L2 acquisition researchers to include other adult learner populations in their studies. For now, we conclude that the combined variable of age and educational context favours university students over child learners in L2 meaning-focused spoken word learning.

2.4.2.2 Treatment

Regarding the incidental learning treatment, the effect size estimates increased in magnitude as expected: task/+interaction > task/–interaction > audiovisual > audio. Nevertheless, only the difference between task/+interaction and task/–interaction was significant, with a small-to-medium effect size. In other words, for L2 spoken word learning, it is beneficial if there is an element of interaction to a learning task. This conforms to earlier literature demonstrating positive effects of interaction (e.g., Keck, Iberri-Shea, Tracy-Ventura & Wa-Mbaleka, 2006; Mackey & Goo, 2007). For future research, it would be interesting to make a distinction within the task/+interaction category between interactive tasks with and without prompted target word production. This has already been done in a few primary studies such as De la Fuente (2002) and Ellis and He (1999). Their findings pointed to a superior role of output. However, the sample sizes in our selection of studies were too small to allow such an investigation.

According to our results, audiovisual treatments had no significant added value over audio-only treatments for L2 word learning. Another observation is that only the difference between task/–interaction and task/+interaction was significant, while the estimated effect of audio (and audiovisual) treatments versus learning task/+interaction was actually larger. This may seem curious, but it is likely a consequence of the fact that task/–interaction versus task/+interaction was often manipulated within studies whereas no studies contrasted audio or audiovisual treatments with treatments involving learning tasks. Thus, the former contrast could be estimated with more precision and therefore more easily achieve significance. Given this consideration, combined with the finding that the effect of the audio versus task/+interaction contrast was actually estimated to be large (albeit non-significant), we are hesitant to confidently conclude that there would be no difference between interactive treatments involving learning tasks on the one hand, and audio and audiovisual treatments on the other. More primary studies are needed from which learning scores for one or more of these treatment types can be extracted so that this question can be reconsidered with more statistical power.

2.4.2.3 Mode of testing

Testing recognition of the newly learned words compared to recall was found to lead to higher effect sizes. Thus, the ability to recognize a word among several alternatives is achieved more easily than the ability to retrieve the word freely from memory. This outcome confirms

Nation's (2001) receptive–productive distinction, which is one of the two dimensions in his model of knowing a word. The other dimension consists of nine subtypes of word knowledge, divided over three domains (form, meaning and use), but for reasons of statistical power these domains and subtypes were not investigated in the current meta-regression. Future studies aiming to evaluate this proposed second dimension of word knowledge can draw from a rich base of primary studies that have already investigated such questions (e.g., Hatami, 2017; Van Zeeland & Schmitt, 2013; Winke et al., 2010).

2.4.2.4 Gain scores

Finally, we investigated two predictors related to the way in which effect sizes were calculated. No significant effect for the use of pre-test to post-test gain scores could be detected, and the effect size was negligible. This is not in accordance with our expectation that studies using a pre-test to post-test design would yield higher effect sizes because the participants already knew what target words to look for and were likely expecting a post-test. Two explanations are conceivable.

First, in some of the studies, explicit efforts had been made to minimise the impact of the pre-test. For example, Nagata, Aline and Ellis (1999) conducted the pre-test three months before the treatment "to ensure that the subjects did not pay focused attention to the lexical items when they performed the task" (p. 140). Montero Perez et al. (2014) told their participants "that such tests are typically administered at the beginning of the academic year" (p. 126). Such approaches are commendable and can help to minimise unwanted pre-test effects.

Second, a large number of the studies that did not use gain scores to control for pre-existing knowledge worked from the assumption that all the target words were unknown to the participants. Although in many cases this assumption seemed justified (e.g., in the studies employing non-words), in some other cases the question of whether this assumption was valid remains. For example, De la Fuente (2002) used words from indigenous languages that are used in Latin American Spanish. It is conceivable that some of her participants, who were students of Spanish at a university in the United States, might have encountered these words while traveling. If in some of the non-gain-scores studies the participants' pre-existing knowledge has been underestimated (thus resulting in larger effect sizes), this may have obscured the comparison between studies that did and did not use gain scores.

The above arguments can explain why, in our sample, there was no effect of gain scores. Of course, this does not mean that future researchers can disregard potential influences of pre-test use. If researchers want to use a pre-test, they should always make efforts to hide the purpose of the pre-test and/or its relationship to a following word learning treatment as well as clearly report when and under what circumstances a pre-test was conducted.

2.4.2.5 Control group

Studies in which treatment groups were compared to true control groups with no exposure to the target words yielded smaller effect sizes than studies that did not contain a control group. The effect was small, but significant. This is in line with our hypothesis, discussed in the Introduction to this chapter. A potential explanation mentioned there was guessing: Participants might (partly) guess words rather than properly learn them, especially when target words are cognates between the L1 and L2. This would then lead to an overestimation of learning effects if no control group (that can also engage in guessing) is used.

Another explanation could be that participants could have learned from the tests themselves, although we tried to exclude this option as much as possible by only including data from repeated post-tests where a test effect seemed unlikely. Regardless of the exact explanation, the finding of a significant control group effect shows the need for control group inclusion in studies that aim to evaluate the effectiveness of word learning interventions. Nation and Webb (2011) argue that control groups can also be useful in determining (and correcting for) unwanted side effects of pre-test use. There is still a lot of room for improvement because, in our sample, only a quarter of the studies included a true control group.

2.4.3 Limitations and recommendations for future research

As meta-analyses and meta-regressions are a product of the primary studies that they are based on, we were dependent on the information reported in the primary studies to run our analyses. Unfortunately, some relevant studies could not be included because means or standard deviations were not reported (see the excluded studies in Appendix S1 in the Online Supporting Information). We urge researchers to always report the means and standard deviations for all treatment and control groups included in their study, a seemingly simple thing to do. Guidelines for reporting quantitative results can be found in Norris, Plonsky, Ross and Schoonen (2015).

Also not often reported in primary studies were correlations between repeated tests on the same participants. Such correlations are needed in meta-analyses and meta-regressions to calculate effect sizes and variances for repeated-measures designs as well as to construct the variance-covariance matrix that is needed to control for dependency between effect sizes. Therefore, we recommend that future researchers report correlations both between pre-test and post-test(s) and between repeated post-tests, or provide the raw data from which these can be calculated. Although our sensitivity analysis showed that our outcomes were not much influenced by borrowing correlations from other studies, it would be better if future meta-analyses could be as precise as possible.

We had originally intended this study to be concerned with incidental learning. However, it was often unclear whether the participants expected to be post-tested on vocabulary (which is one important definition of incidental learning), and we therefore resorted to studying meaning-focused learning. To enable future researchers to properly study incidental learning (and perhaps contrast it with intentional learning), it is important that all authors

of primary studies report whether the post-test was announced to the participants. Ideally, after finishing the experiment, they would also interview the participants about whether they were expecting a post-test. If researchers wish to study incidental learning and have the post-test come as a surprise, they should try to prevent situations where participants can guess that a post-test might be administered.

In the meta-regression, we focused on three substantive variables (age, treatment, and mode of testing) that are central to L2 word learning. Many other variables of potentially high importance could not be included. For example, one variable that can influence L2 word learning is exposure frequency (e.g., Brown et al., 2008; Van Zeeland & Schmitt, 2013; although exposure effects are not always found, as in Gullberg et al., 2012). However, it was not possible to control frequency of exposure in this study (or make it a predictor), because this variable was either not controlled in some of the primary studies (e.g., De la Fuente, 2002; Donkaewbua, 2009) or no information about exposure frequency was provided (e.g., Koolstra & Beentjes, 1999; Medina, 1990). Similarly, there was no possibility of incorporating retention interval as a variable in the current study. Proficiency is another important variable that could not be included for various reasons, one of them being the sample size of our data set. In addition, proficiency was measured with different instruments (or was not measured at all) across studies, and within studies the participants sometimes were of different proficiency levels. We recommend that future primary researchers control and report exposure frequency, retention interval and proficiency, and that future meta-analysts include these variables in their designs.

2.5 SUMMARY AND CONCLUSIONS

Until now, no consensus has existed on the effectiveness of word learning from meaning-focused exposure to spoken L2 input. Our meta-analysis showed that this type of learning is very well possible, on average yielding a large effect. Whether this finding also applies to studies in which the learning was incidental also according to other definitions is still to be seen. For this to be investigated, primary studies should first become more transparent in reporting the exact pre-test and post-test instructions that the participants were given and ideally verify post-test expectancy through interviews.

In our meta-regression, we detected significant effects for age (higher scores for older participants), treatment (higher scores for learning tasks with interaction than without), and mode of testing (higher scores on recognition tests than on recall tests). Studies using a true control group yielded lower (probably more realistic) effect sizes, but we detected no difference between studies that did and did not make use of gain scores. All of these novel insights could be extracted from already existing research, which shows the great potential that the technique of meta-regression has for furthering our knowledge in any given domain of language learning.

APPENDIX A: CALCULATING META-ANALYTIC STATISTICS

In this appendix, we detail how the meta-analytic statistics described in the Methods section of the chapter (2.2.4) were calculated for the four different types of study designs. All 105 effect sizes and associated variances can be found in Appendix S2 in the Online Supporting Information, which also shows the characteristics associated with each effect size (sample size, age group, etc.). We will begin with some general remarks that apply to all four study design types.

Averaging within tests

Sometimes, in the primary studies the test results were broken down by variables that are not of interest in the current study. For example, in the second experiment by Gullberg et al. (2012) some items were presented with gestural highlighting, while others were not. Means and standard deviations were reported separately for each presentation type. In these cases, we averaged over those outcomes that could not be distinguished in terms of our five predictors. Standard deviations were pooled according to the following formula:

$$SD_{\text{pooled}} = \sqrt{\frac{{SD_1}^2+{SD_2}^2+\cdots+{SD_k}^2}{k}} \qquad \text{(Statistics How To, 2017)}$$

In case of unequal sample sizes, we calculated the weighted average and pooled standard deviation. The latter was calculated as follows:

$$SD_{\text{pooled}} = \sqrt{\frac{(n_1-1){SD_1}^2+(n_2-1){SD_2}^2+\cdots+(n_k-1){SD_k}^2}{n_1+n_2+\cdots+n_k-k}} \qquad \text{(Statistics How To, 2017)}$$

We only averaged the results on vocabulary items within tests, not between. Tests conducted at different moments in time or with different testing instruments reflect different kinds of knowledge, which we did not want to lose by averaging. As described in the Analysis section (2.2.6), we can deal with the dependency that exists between effect sizes coming from the same study by including the covariance of these measures in the analysis. We will now discuss how learning scores were calculated for the four types of study designs.

1. Comparison within treatment group(s): Post-test scores to fixed baseline

For 14 studies (yielding 65 effect sizes), we compared the post-test scores to a baseline. This means that the same level of pre-existing knowledge of the target items was assumed for all participants, and that the post-test scores were compared to this assumed baseline. The baseline for each primary study was calculated by the author of this thesis.

If there was reason to believe that all target items were unknown to the participants and the study employed open questions, the baseline for comparison was set to 0 (e.g., for De la Fuente, 2002, who used words from indigenous languages in Latin-American countries). If no pre-existing knowledge was assumed, but the test consisted of multiple-choice questions, the baseline was set according to the probability of correct guessing. For example, Gullberg

et al. (2012) presented yes/no questions, which makes the probability of a correct guess 50%. Since their scores were presented as percentages, the baseline was set to 50. Slightly more complicated are Brown et al. (2008) and Van Zeeland and Schmitt (2013), who both offered an "I don't know" answer option along with the other answers. Because guessing was hereby discouraged, we set the baseline to 0. Of course, it is still possible that the participants made a guess anyway. For a more conservative effect size estimate, the baseline could be set to 0.25 * 28 = 7 (for the Brown et al., 2008, example).

In some studies of the 'post-test to fixed baseline' type, the primary researchers still conducted a pre-test. If they did, however, this pre-test was used to select the least-known/unknown target items for the study, rather than calculate pre-test to post-test gain scores for each participant individually. For example, in the Tokyo study in Ellis et al. (1994), 19 items were selected which were unknown to at least 88% of the students. Subsequently, the treatment and the post-test were conducted on the final set of 19 target words. We calculated the baseline for a recall translation task as (1 - 0.88) * 19 = 2.28.

It should be noted that since all target items were unknown to at least 88% of the students (but potentially to more), this baseline of 2.28 represents a maximum estimate of students' potential pre-existing knowledge. The resulting effect size thus is likely to be more conservative than effect sizes calculated against a baseline that is derived from an estimation of participants' average pre-existing knowledge (such as in Brown et al., 2008). However, as the participants' pre-test scores were not available for individual items, this was the only possible approach.

The one-sample ('os') effect size statistics were calculated as follows:

$$d_{\text{os}} = \frac{M_{\text{sample}} - \text{baseline}}{SD_{\text{sample}}}$$ (Cohen, 1988)

$$J_{\text{os}} = 1 - \frac{3}{4(N-2)-1}$$ (Hedges & Olkin, 1985)

$$g_{\text{os}} = d_{\text{os}} * J_{\text{os}}$$ (Borenstein, 2009, p. 226)

Regarding the sampling variance of d_{os}, no widely-used formula for computing the variance for data from one-sample studies was available. A simulation comparing three different formulas for computing this variance showed that the standard formula for v_d with r set to 0.5 yielded the least-biased estimates (Borenstein, 2009; Koenig, Eagly, Mitchell & Ristikari, 2011; Sampling variance for meta-analysis one-sample data, 2016). Thus, we used that approach:

$$v_{os} = J_{\text{os}}^2 * \left(\frac{1}{n} + \frac{d^2}{2n}\right) * 2(1-r)$$

2. Comparison within treatment group(s): Pre-test to post-test gain scores

Eleven studies (yielding 23 effect sizes) in the sample had a repeated-measures design without true control group, using a pre-test and one or more post-tests. Eight out of these studies reported means and standard deviations for the pre-test and post-test, whereas three studies only reported the means and standard deviations of the gain scores (see Appendix S2 in the Online Supporting Information). In the latter scenario, d_{rm} for repeated measures can be calculated as follows:

$$d_{\mathrm{rm}} = \frac{M_{\mathrm{gain}}}{SD_{\mathrm{gain}}} \qquad \text{(Morris \& DeShon, 2002, p. 107)}$$

However, Morris and DeShon (2002) point out that a standard deviation of gain scores generally is not the same as a (pooled) standard deviation of raw scores, such as would be used for calculating effect sizes in an independent-group ('ig') design (see below, Comparison 3). Thus, effect sizes calculated from these two types of standard deviations are not comparable. Fortunately, there is a transformation to solve this issue (*r* represents the correlation between pre-test and post-test scores):

$$d_{\mathrm{ig}} = d_{\mathrm{rm}} * \sqrt{2 * (1 - r)} \qquad \text{(Morris \& DeShon, 2002, p. 111)}$$

However, Lakens (2013) argues that the above effect size can sometimes be "unreasonably conservative" when correlations between observations are high (Lakens, 2013, p. 5). He recommends simply using the average standard deviation of both measures as a standardiser:

$$d_{\mathrm{av}} = \frac{M_{\mathrm{post}} - M_{\mathrm{pre}}}{\left(\frac{SD_{\mathrm{pre}} + SD_{\mathrm{post}}}{2}\right)} \qquad \text{(Lakens, 2013, p. 5)}$$

Therefore, we calculated d_{av} wherever possible. In two cases where pre-test standard deviations were not provided (Hsu, Hwang, Chang & Chang, 2013; Yeung et al. 2016), we divided the average gain score by the post-test standard deviation only.

For the three studies that reported only gain scores we calculated d_{rm} and transformed it into d_{ig}. A complication in this was that none of these three studies reported the correlation between pre-test and post-test scores. This is a well-known problem in meta-analysis, but it can be dealt with by using data from other sources to estimate this correlation (Borenstein, 2009, p. 227; see also the main Analysis section). Therefore, we selected correlation coefficients from other studies in the data set that came closest to the study with missing correlations in terms of design, treatment, etc. Appendix S2 in the Online Supporting Information shows which correlations from which studies were taken as substitutes (tab: Effect sizes, column: Borrowed correlation). Appendix B reports a sensitivity analysis in which we investigated how different assumptions for missing correlations impacted the meta-analytic outcomes.

g_{ig} and g_{av} for all studies were obtained by multiplying d_{ig} and d_{av} with J_{rm}:

$$J_{\mathrm{rm}} = 1 - \frac{3}{4(N-2)-1}$$ (Hedges & Olkin, 1985)

$$g_{\mathrm{ig/av}} = d_{\mathrm{ig/av}} * J_{\mathrm{rm}}$$ (Borenstein, 2009, p. 226)

The sampling variance for effect sizes coming from pre-test to post-test data is given as follows:

$$v_d = J^2 * \left(\frac{1}{n} + \frac{d^2}{2n}\right) * 2(1-r)$$ (Borenstein, 2009, p. 227)

Again, if no correlation coefficient was available, it was taken from another, comparable study, and this was marked in Appendix S2 in the Online Supporting Information.

3. Treatment group(s) compared to true control group: Post-test scores only

Four studies (yielding 11 effect sizes) compared post-test scores of a treatment group to post-test scores of a control group. Effect sizes for independent groups ('ig') and their sampling variances were calculated as follows (T = treatment, C = control):

$$d_{\mathrm{ig}} = \frac{M_{\mathrm{T}} - M_{\mathrm{C}}}{SD_{\mathrm{within}}}$$ (Borenstein, 2009, p. 226)

$$SD_{\mathrm{within}} = \sqrt{\frac{(n_{\mathrm{T}}-1)*SD_{\mathrm{T}}^2 + (n_{\mathrm{C}}-1)*SD_{\mathrm{C}}^2}{n_{\mathrm{T}}+n_{\mathrm{C}}-2}}$$ (Borenstein, 2009, p. 226)

$$J_{\mathrm{ig}} = 1 - \frac{3}{4*df-1}$$ (Borenstein, 2009, p. 226)

$$g_{\mathrm{ig}} = d_{\mathrm{ig}} * J_{\mathrm{ig}}$$ (Borenstein, 2009, p. 226)

$$v_{\mathrm{ig}} = J_{\mathrm{ig}}^2 * \frac{n_{\mathrm{T}}+n_{\mathrm{C}}}{n_{\mathrm{T}}*n_{\mathrm{C}}} + \frac{d^2}{2(n_{\mathrm{T}}+n_{\mathrm{C}})}$$ (Borenstein, 2009, p. 226)

4. Treatment group(s) compared to true control group: Pre-test to post-test gain scores

Four studies (yielding six effect sizes) presented data from a pre-test to post-test design with a control group. Morris (2008) investigates three effect size estimates for such designs. He recommends an effect size based on the mean pre-test to post-test gain in the treatment group minus the mean pre-test to post-test gain in the true control group, divided by the pooled pre-test standard deviation.

However, applying this approach to the data, the Vidal (2011) study yielded effect sizes that were extreme outliers (g = 11.42 and g = 5.62). This can be explained by the fact that the pre-test standard deviations in Vidal (2011) are much smaller than the post-test standard deviations, but the latter are not taken into account. In addition to producing these outliers, not taking the post-test standard deviation into account is inconsistent with the way effect sizes for studies without a control group were calculated. Therefore, even though disfavoured by Morris (2008), we used one of his alternative effect size estimates that pools the standard deviation across both pre-test and post-test measurements (for both the treatment and control condition). Statistics for the pre-test-to-post-test-control design ('ppc') can be calculated as follows (T = treatment, C = control):

$$d_{\text{ppc}} = \frac{(M_{\text{post,T}} - M_{\text{pre,T}}) - (M_{\text{post,C}} - M_{\text{pre,C}})}{SD_{\text{pre+post}}}$$

(Morris, 2008, p. 370)

$$SD_{\text{pre+post}} = \sqrt{\frac{(n_{\text{T}}-1)*SD^2_{\text{pre,T}} + (n_{\text{C}}-1)*SD^2_{\text{pre,C}} + (n_{\text{T}}-1)*SD^2_{\text{post,T}} + (n_{\text{C}}-1)*SD^2_{\text{post,C}}}{2(n_{\text{T}}+n_{\text{C}}-2)}}$$

(Morris, 2008, p. 370)

As usual, d_{ppc} was then multiplied with a correction factor J_{ppc} (called c_{pp} by Morris, 2008) to address small-sample bias:

$$J_{\text{ppc}} = 1 - \frac{3}{4(2*n_{\text{T}} + 2*n_{\text{C}} - 4) - 1}$$

(Morris, 2008, p. 370)

$$g_{\text{ppc}} = d_{\text{ppc}} * J_{\text{ppc}}$$

(Morris, 2008, p. 370)

An issue with the effect size d_{ppc} is that its exact sampling variance is currently unknown. However, it is expected to be smaller than the other alternatives discussed in Morris (2008), and to approach the sampling variance of the above-discussed alternative based on pre-test standard deviation only, as the correlation between pre-test and post-test scores approaches 1. We calculated the variance using the formula for the alternative effect size based on pre-test standard deviations only. If this formula is indeed expected to overestimate the true variance,

this would mean that the two studies of this type are weighted less in this meta-analysis than they would otherwise. This is unfortunate, but still seems to be the better option compared to weighing them too much in the analysis. Thus, v_{ppc} was calculated as follows:

$$v_{\mathrm{ppc}} = J^2_{\mathrm{ppc}} * \ 2 * (1 - r) * \frac{n_{\mathrm{T}} + n_{\mathrm{C}}}{n_{\mathrm{T}} * n_{\mathrm{C}}} * \frac{n_{\mathrm{T}} + n_{\mathrm{C}} - 2}{n_{\mathrm{T}} * n_{\mathrm{C}} - 4} * \left(1 + \frac{d^2_{\mathrm{ppc}}}{2(1-r) * \left(\frac{n_{\mathrm{T}} + n_{\mathrm{C}}}{n_{\mathrm{T}} * n_{\mathrm{C}}}\right)}\right) - d^2_{\mathrm{ppc}}$$

(Morris, 2008, p. 370)

APPENDIX B: SUPPLEMENTS TO THE ANALYSIS

When no effect sizes were excluded based on their Cook's distance value, the following model estimates were obtained (Table A).

Table A. Results from Model 2, as computed on the full data set.

Fixed effects	β	*SE*	*z*	*p*	Lower bound	Upper bound
Intercept	0.32	0.46	0.70	.49	-0.58	1.23
Age: high school	0.48	0.54	0.90	.37	-0.57	1.53
Age: university	0.71	0.36	1.98	**.048**	0.01	1.41
Treatment: audiovisual	-0.05	0.22	-0.22	.83	-0.48	0.38
Treatment: task/-interaction	0.01	0.43	0.03	.98	-0.84	0.86
Treatment: task/+interaction	0.66	0.46	1.45	.15	-0.23	1.55
Testing: recognition	0.38	0.14	2.76	**< .01**	0.11	0.66
Gain scores: yes	0.08	0.31	0.25	.80	-0.53	0.68
Control group: yes	-0.55	0.29	-1.89	.06	-1.12	0.02
Random effects	**Variance**	***SD***				
Intercept (study)	0.36	0.60				
Intercept (sample)	0.19	0.43				

Note. k = 105. The intercept represents the following combination of variable levels: Age = kindergarten/elementary school, Treatment = audio, Testing = recall, Gain scores = no, and Control group = no. Significant p-values are printed in bold.

Profile likelihood plots

Profile likelihood plots of the variance components are shown in Figure A (Model 1) and Figure B (Model 2, three cases with $D_i > 0.85$ excluded). In this procedure, σ_1^2 (variance at the study level) and σ_2^2 (variance at the sample level) were fixed at different values (i.e., at all the positions of the dots on the x-axis). For each value of σ_1^2 and σ_2^2, the (logarithm of the) likelihood over the remaining model parameters, such as the fixed effects, was estimated (Viechtbauer, 2010). This means it was estimated how likely the values of these parameters were given the observed data. Less negative values of the logarithm indicate a higher likelihood than more negative values. It can be seen that the likelihood is estimated to be the highest for the values of σ_1^2 and σ_2^2 that had been estimated in the original models (0.31 and 0.21 for Model 1, and 0.49 and 0.05 for Model 2). This, and also the fact that the log-likelihoods become more negative as the values of σ^2 move away from the parameter estimates, suggests that we can be "fairly confident" that our meta-analytic models could identify the variance components (Viechtbauer, 2017).

Figure A. Profile likelihood plots for the variance components of Model 1.

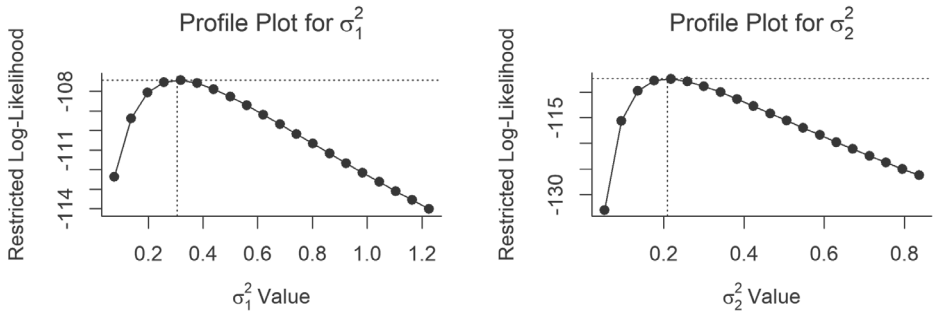
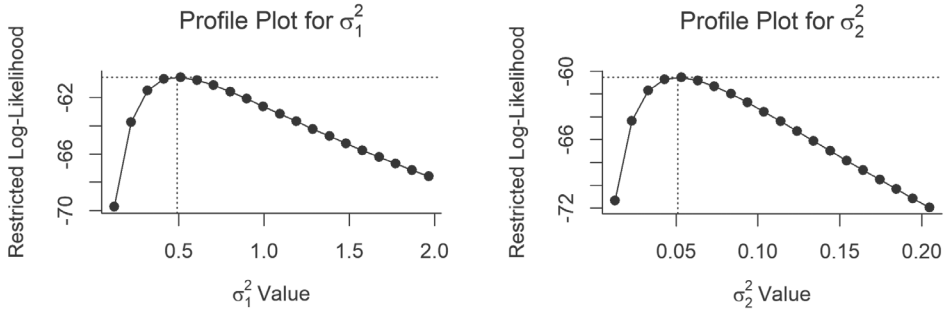


Figure B. Profile likelihood plots for the variance components of Model 2.



Sensitivity analysis

For some of the studies in our sample, no correlation coefficients were reported. Correlation coefficients are needed in meta-analysis for multiple purposes: to calculate effect sizes and variances for repeated-measures designs, and to control for the dependence between independent samples in the same study (as effect sizes coming from the same lab or author can be expected to be more alike than effect sizes coming from different labs or authors). As described in the main Analysis section and in Appendix A, we used correlation coefficients from comparable studies in such cases to estimate the missing values.

Here, we report the outcomes of two sensitivity analyses we conducted to investigate how robust the outcomes of our study were to variation in the assumed correlation coefficients. To this end, we ran Model 1 and Model 2 twice more. First, we subtracted 0.10 from all assumed correlation coefficients, and then we added 0.10 to all assumed correlation coefficients. Any resulting correlations coming out as larger than 1 were set to 1. Table B shows the outcomes for Model 1: The data are not affected much by changes in correlation of ±0.10.

Table B. Model 1 rerun on two data sets in which assumed correlations were increased and decreased by 0.10.

	β	*SE*	*z*	*p*	Lower bound	Upper bound	σ_1^2	σ_2^2
Original data set	1.05	0.12	8.77	**< .001**	0.81	1.28	0.31	0.21
All correlations – 0.10	1.07	0.12	9.00	**< .001**	0.84	1.31	0.30	0.22
All correlations + 0.10	1.01	0.12	8.15	**< .001**	0.77	1.25	0.35	0.20

Note. k = 105. Significant *p*-values are printed in bold.

Reruns of Model 2 on the adjusted data sets are shown in Tables C and D. For comparison purposes, Table 2 from the Results section is reprinted here as well.

Table 2. Results from Model 2 after the exclusion of three influential cases.

Fixed effects	β	*SE*	*z*	*p*	Lower bound	Upper bound
Intercept	0.00	0.49	-0.01	.99	-0.96	0.95
Age: high school	0.74	0.59	1.27	.21	-0.41	1.89
Age: university	0.92	0.41	2.26	**.02**	0.12	1.72
Treatment: audiovisual	0.07	0.16	0.43	.67	-0.25	0.39
Treatment: task/-interaction	0.10	0.44	0.22	.83	-0.76	0.95
Treatment: task/+interaction	0.73	0.45	1.61	.11	-0.16	1.61
Testing: recognition	0.42	0.09	4.42	**< .001**	0.23	0.60
Gain scores: yes	0.03	0.32	0.11	.91	-0.59	0.66
Control group: yes	-0.47	0.23	-2.03	**.04**	-0.93	-0.02
Random effects	**Variance**	***SD***				
Intercept (study)	0.49	0.70				
Intercept (sample)	0.05	0.23				

Note. k = 102. The intercept represents the following combination of variable levels: Age = kindergarten/elementary school, Treatment = audio, Testing = recall, Gain scores = no, and Control group = no. Significant *p*-values are printed in bold.

Table C. Results from Model 2, when assumed correlations were decreased by 0.10.

Fixed effects	β	*SE*	*z*	*p*	Lower bound	Upper bound
Intercept	-0.08	0.49	-0.16	.87	-1.04	0.88
Age: high school	0.74	0.58	1.27	.20	-0.40	1.88
Age: university	0.95	0.41	2.34	**.02**	0.15	1.74
Treatment: audiovisual	0.13	0.18	0.76	.45	-0.21	0.48
Treatment: task/-interaction	0.15	0.44	0.34	.73	-0.70	1.00
Treatment: task/+interaction	0.80	0.45	1.78	.07	-0.08	1.68

Testing: recognition	0.44	0.10	4.37	**< .001**	0.24	0.64
Gain scores: yes	0.07	0.32	0.20	.84	-0.56	0.69
Control group: yes	-0.47	0.24	-1.94	.052	-0.95	0.00
Random effects	**Variance**	***SD***				
Intercept (study)	0.47	0.69				
Intercept (sample)	0.06	0.25				

Note. The intercept represents the following combination of variable levels: Age = kindergarten/elementary school, Treatment = audio, Testing = recall, Gain scores = no, and Control group = no. *k* = 102. Significant *p*-values are printed in bold.

Table D. Results from Model 2, when assumed correlations were increased by 0.10.

Fixed effects	**β**	***SE***	***z***	***p***	**Lower bound**	**Upper bound**
Intercept	0.04	0.47	0.08	.94	-0.89	0.97
Age: high school	0.73	0.60	1.23	.22	-0.44	1.90
Age: university	0.85	0.41	2.07	**.04**	0.05	1.66
Treatment: audiovisual	0.09	0.13	0.71	.48	-0.16	0.34
Treatment: task/-interaction	0.11	0.43	0.27	.79	-0.73	0.95
Treatment: task/+interaction	0.70	0.44	1.60	.11	-0.16	1.57
Testing: recognition	0.35	0.08	4.29	**< .001**	0.19	0.51
Gain scores: yes	-0.04	0.32	-0.11	.91	-0.67	0.59
Control group: yes	-0.42	0.20	-2.13	**.03**	-0.81	-0.03
Random effects	**Variance**	***SD***				
Intercept (study)	0.52	0.72				
Intercept (sample)	0.03	0.18				

Note. k = 102. The intercept represents the following combination of variable levels: Age = kindergarten/elementary school, Treatment = audio, Testing = recall, Gain scores = no, and Control group = no. Significant *p*-values are printed in bold.

Again, we see that the differences in outcomes are minimal, both with regard to beta estimates and with regard to significance. Thus, from this sensitivity analysis we conclude that our meta-analysis and meta-regression were robust against variation in correlation coefficients.

3.

Interactive L2 vocabulary acquisition in a lab-based immersion setting

This chapter is based on:

De Vos, J. F., Schriefers, H., Ten Bosch, L., & Lemhöfer, K. (2019). *Interactive L2 vocabulary acquisition in a lab-based immersion setting.* Manuscript accepted for publication.

ABSTRACT

We investigated to what extent second language (L2) word learning in spoken interaction takes place when learners are unaware of taking part in a language learning study. Using a novel paradigm, 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. This study provides a new and effective paradigm for approximating naturalistic, interactive L2 vocabulary learning in the lab.

3.1 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.

3.1.1 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 and 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 was given in Chapter 2. 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 Chapter 2, and 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. 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 in Chapter 4[1], where we 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 Chapter 4 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.

3.1.2 The present study

The present study used a similar experimental set-up as the study described in Chapter 4, but was different in the fact that the current participants 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

[1] The study described in Chapter 4 was conducted after the study described here in Chapter 3, but was published before. Therefore, while the study in Chapter 4 builds on the paradigm we developed in the current study, in this chapter I sometimes already refer to Chapter 4. This approach follows the article version of Chapter 3 (the revised version of which has been submitted), in which we also refer to the published article version of Chapter 4.

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 object 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 object 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).

3.1.2.1 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 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.

3.1.2.2 Exposure frequency

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, 2014b; 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. (2014b) 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.

3.1.2.3 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 above studies, 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.

3.1.2.4 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).

3.1.2.5 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, 2000; 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).

3.1.3 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?

3.2 METHODS

3.2.1 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, 3.2.3.2). 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 study described in Chapter 4 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.

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

	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 the Methods section (3.2.2.2). Variables marked with an asterisk were self-rated on a 1-5 Likert scale.

3.2.2 Materials

3.2.2.1 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. 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. A list of all items can be found in Appendix A at the end of this chapter. As mentioned in the Introduction to this chapter, the items (both targets and fillers) were selected on an individual basis for each participant. This means that from the final item pool, a different subset was extracted for each participant. This will be discussed in more detail in the Procedure section (3.2.3.2).

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 program (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, 2008). 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.

3.2.2.2 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 Appendix B. 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).

3.2.3 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.

3.2.3.1 Sorting task (the pre-test)

The experimenter (a female native speaker of Dutch and the author of this thesis) 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.

3.2.3.2 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.

3.2.3.3 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 15x15 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" (in Dutch). 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.

3.2.3.4 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.

3.2.3.5 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.

3.2.3.6 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.

3.2.4 Analysis

3.2.4.1 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 statistical model, we left these measures out of the main statistical analysis. However, explorative correlations are reported in Appendix H.[2]

[2] 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.

3.2.4.2 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 of 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.

3.2.4.3 Scoring

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 level 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 section 4.2.4.3 in Chapter 4.

3.2.4.4 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 E 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 the main part of this chapter 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 D; the final models are presented in the Results section (3.3.3).

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.

3.3 RESULTS

3.3.1 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.

3.3.2 Descriptive statistics

The learning scores are depicted graphically in Figure 1. Pre-test scores were at 0 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 C in Appendix C contains descriptives for all subcombinations of predictor levels as well. As explained in the Methods section (3.2.4.3), in both the figure and the tables the dependent variable is the average percentage of correctly produced phonemes per target word utterance.

Figure 1. Mean scores across the four testing moments (EF = Exposure frequency). Error bars represent 95% confidence intervals based on a bootstrap.

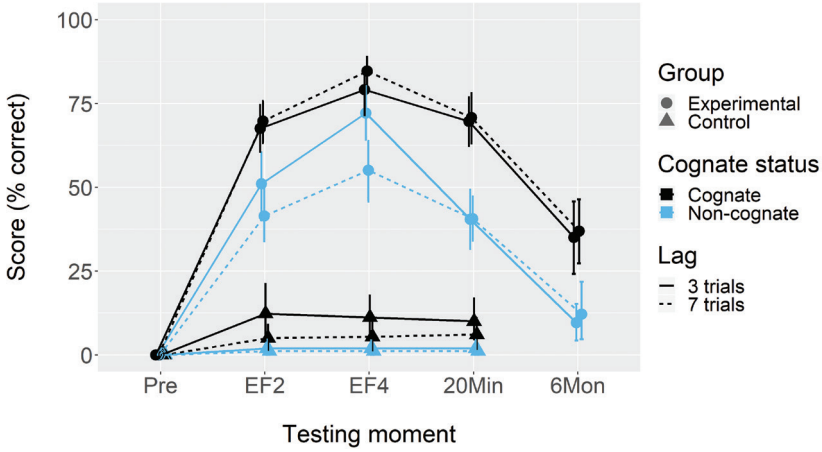


Table 2. Percentage of correctly produced phonemes per target word during the price comparison task (i.e., the learning phase).

		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 minutes	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**

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 0, 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, where non-cognates seem to benefit from having a shorter lag of three trials, but this is not the case for cognates, where if anything the effect is reversed. We had no hypothesis about the presence of such an interaction, and explored it further in Appendix E. Appendices H and I contain additional analyses at the participant and item level.

3.3.3 Model comparisons

Appendix D 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 Appendix D. The final retention model was: (Number of correct phonemes, Number of incorrect phonemes) ~ 1 + Cognate status + Retention interval + Lag + (1 + Cognate status * Retention interval + Lag | Participant) + (1 + Retention interval * Lag | Word). In section 3.3.4.4, we will describe how the final models' fit to the data was evaluated.

3.3.4 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.

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.

3.3.4.1 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 Appendix E, 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 0, indicate large effects. The exact interpretation of ORs, as well as the interpretation of logit estimates, is explained in more detail in Appendix F.

3.3.4.2 Outcomes of the learning phase

As can be seen from the main Group effect in Table 4, the participants in the experimental group significantly outperformed the participants in the control group. This 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 E showed that the interaction between Group, Cognate status and Lag during the learning phase that seemed visible in Figure 1 did not reach significance.

3.3.4.3 Long-term outcomes

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

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 minutes		-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 minutes	0.64	0.80			
	RI = 6 months	5.19	2.28			
	L = 7 trials	0.91	0.95			
	CS = Non-cognate : RI = 20 minutes	0.69	0.83			
	CS = Non-cognate : RI = 6 months	21.67	4.66			
Word	(Intercept)	13.79	3.71			
	RI = 20 minutes	2.75	1.66			
	RI = 6 months	23.14	4.81			
	L = 7 trials	5.74	2.40			
	RI = 20 minutes : 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.

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% ($\beta = -4.70$, $OR = 0.01$, $z = -7.67$, $p < .001$); 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 scores were still significantly above 0.

The explorative retention model presented in Appendix E 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.

3.3.4.4 The 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 Appendix G. 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 0. Furthermore, 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 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 G, 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.

3.4 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 (3.3.1), 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.

3.4.1 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/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 further explored in the next chapter.

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 Chapter 4, the design of which was less interactive than the design of the current study. In Chapter 4, the 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 scores in the current study were higher than in the study described in Chapter 4: 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, in Chapter 4 we show 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.

3.4.2 Predictors of L2 word learning

3.4.2.1 Cognate status

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), 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 result 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 0 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 0 is in line with the meta-regression described in Chapter 2, 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.

3.4.2.2 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. (2014b). 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. (2014b, 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., 2014b, 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 the learners. As Bisson et al. (2014b, 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.

3.4.2.3 Lag

No 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 Appendix E). 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.

3.4.2.4 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 0. In calculating these scores, 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.

3.4.3 Relation to the immersion and incidental learning literature

At the beginning of this chapter, 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 chapter, 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 the General discussion to this thesis (section 7.3.1), I discuss studies in other domains of second language acquisition in which the effects of awareness of a study's aim, or of an upcoming post-test, were investigated.

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 to this chapter. 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.

3.5 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 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. 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.

APPENDIX A: ITEMS

Table A contains all the target items that were used in this experiment, Table B contains the fillers.

Table 1. Target items.

Dutch	German	English	Category	Cognate status	Length (phonemes)
emmer	Eimer	bucket	children	cognate	4
gum	Radiergummi	eraser	children	cognate	3
knuffel	Kuscheltier	cuddly toy	children	cognate	6
peddel	Paddel	paddle	children	cognate	5
stelt	Stelze	stilt	children	cognate	5
stokpaard	Steckenpferd	hobby horse	children	cognate	8
tamboerijn	Tamburin	tambourine	children	cognate	8
toverstaf	Zauberstab	magic wand	children	cognate	9
wip	Wippe	seesaw	children	cognate	3
zwemvleugel	Schwimmflügel	water wing	children	cognate	10
knikker	Murmel	marble	children	non-cognate	6
luier	Windel	nappy	children	non-cognate	5
puntenslijper	Anspitzer	pencil sharpener	children	non-cognate	11
rammelaar	Rassel	rattle	children	non-cognate	7
romper	Body	onesie	children	non-cognate	6
sambabal	Rassel	maraca	children	non-cognate	8
skelter	Kettcar	go-kart	children	non-cognate	7
slinger	Girlande	bunting banner	children	non-cognate	6
speen	Schnuller	teat	children	non-cognate	4
tol	Kreisel	top	children	non-cognate	3
keppel	Kippa	yarmulka	clothes	cognate	5
kraag	Kragen	collar	clothes	cognate	4
kroon	Krone	crown	clothes	cognate	4
mijter	Mitra	mitre	clothes	cognate	5
pruik	Perücke	wig	clothes	cognate	4
reddingsvest	Rettungsweste	life jacket	clothes	cognate	10
rits	Reißverschluss	zipper	clothes	cognate	4
schort	Schürze	apron	clothes	cognate	5
staf	Stab	staff	clothes	cognate	4
tulband	Turban	turban	clothes	cognate	7

gesp	Gürtelschnalle	clasp	clothes	non-cognate	4
hes	Warnveste	smock	clothes	non-cognate	3
kous	Strumpf	stocking	clothes	non-cognate	3
pet	Kappe	cap	clothes	non-cognate	3
slab	Lätzchen	bib	clothes	non-cognate	4
speldje	Spange	hair clip	clothes	non-cognate	7
tooi	Federschmuck	headdress	clothes	non-cognate	3
tuinbroek	Latzhose	dungarees	clothes	non-cognate	7
veter	Schnürsenkel	shoelace	clothes	non-cognate	5
waaier	Fächer	fan	clothes	non-cognate	5
bezem	Besen	broom	household	cognate	5
citruspers	Zitruspresse	lemon squeezer	household	cognate	10
kist	Kiste	chest	household	cognate	4
kurk	Korken	cork	household	cognate	4
lantaarn	Laterne	lantern	household	cognate	7
onderzetter	Untersetzer	coaster	household	cognate	10
servet	Serviette	serviette	household	cognate	6
slang	Schlauch	(garden) hose	household	cognate	4
stamper	Stampfer	stamp(er)	household	cognate	7
wafelijzer	Waffeleisen	waffle iron	household	cognate	9
broodrooster	Toaster	toaster	household	non-cognate	10
dienblad	Serviertablett	tray	household	non-cognate	7
dweil	Wischmopp	mop	household	non-cognate	4
garde	Schneebesen	whisk	household	non-cognate	5
kapstok	Kleiderständer	coathooks	household	non-cognate	7
kooi	Käfig	cage	household	non-cognate	3
lessenaar	Pult	music stand	household	non-cognate	7
rietje	Strohhalm	straw	household	non-cognate	5
stolp	Glasglocke	(bell-)glass	household	non-cognate	5
vergiet	Sieb	colander	household	non-cognate	6
aambeeld	Amboss	anvil	tools	cognate	6
gieter	Gießkanne	watering can	tools	cognate	5
heggenschaar	Heckenschere	hedge-clippers	tools	cognate	8
hengel	Angel	fishing rod	tools	cognate	5
klapper	Filmklappe	clapper	tools	cognate	6
kruk	Krücke	crutch	tools	cognate	4
schroef	Schraube	screw	tools	cognate	5
vijl	Feile	file	tools	cognate	3

zaag	Säge	saw	tools	cognate	3
zeis	Sense	scythe	tools	cognate	3
brandblusser	Feuerlöscher	fire extinguisher	tools	non-cognate	11
buis	Rohr	tube	tools	non-cognate	3
dobber	Schwimmer	float	tools	non-cognate	5
klos	Rolle	reel (of cotton)	tools	non-cognate	4
kruiwagen	Schubkarre	wheelbarrow	tools	non-cognate	7
kwast	Pinsel	brush	tools	non-cognate	5
passer	Zirkel	compass	tools	non-cognate	5
spijker	Nagel	nail	tools	non-cognate	6
spuit	Spritze	syringe	tools	non-cognate	4
vijzel	Mörser	mortar	tools	non-cognate	5

Table B. Filler items.

Dutch	German	English	Category
bal	Ball	ball	children
ballon	Ballon	balloon	children
banaan	Banane	banana	children
boek	Buch	book	children
fluit	Flöte	flute	children
frisbee	Frisbee	frisbee	children
gameboy	Gameboy	game boy	children
gitaar	Gitarre	guitar	children
hond	Hund	dog	children
kat	Katze	cat	children
koe	Kuh	cow	children
konijn	Kaninchen	rabbit	children
muffin	Muffin	muffin	children
paard	Pferd	horse	children
piano	Klavier	piano	children
pleister	Pflaster	plaster	children
pop	Puppe	doll	children
postzegel	Briefmarke	stamp	children
rugzak	Rucksack	rucksack	children
skateboard	Skateboard	skateboard	children
tandenborstel	Zahnbürste	toothbrush	children
vis	Fisch	fish	children
vogel	Vogel	bird	children

wekker	Wecker	alarm	children
armband	Armband	bracelet	clothes
beha	BH	bra	clothes
bikini	Bikini	bikini	clothes
blouse	Bluse	blouse	clothes
bril	Brille	glasses	clothes
broek	Hose	trousers	clothes
handdoek	Handtuch	towel	clothes
handschoen	Handschuh	glove	clothes
jas	Mantel	coat	clothes
koffer	Koffer	suitcase	clothes
muts	Mütze	hat	clothes
pak	Anzug	suit	clothes
paraplu	Regenschirm	umbrella	clothes
parfum	Parfum	perfume	clothes
pyjama	Pyjama	pyjamas	clothes
ring	Ring	ring	clothes
rok	Rock	skirt	clothes
schoen	Schuh	shoe	clothes
sjaal	Schal	scarf	clothes
sok	Socke	sock	clothes
tas	Tasche	bag	clothes
trui	Pullover	jumper	clothes
t-shirt	T-Shirt	t-shirt	clothes
zonnebril	Sonnenbrille	sunglasses	clothes
lippenstift	Lippenstift	lipstick	clothes
bank	Sofa	sofa	household
bed	Bett	bed	household
bord	Teller	plate	household
deksel	Deckel	lid	household
deur	Tür	door	household
douche	Dusche	shower	household
glas	Glas	glass	household
gordijn	Gardine	curtain	household
kaars	Kerze	candle	household
kam	Kamm	comb	household
kast	Schrank	closet	household
klok	Uhr	clock	household

koelkast	Kühlschrank	fridge	household
kopje	Tasse	cup	household
kussen	Kissen	pillow	household
lamp	Lampe	lamp	household
matras	Matratze	mattress	household
pan	Pfanne	pan	household
plant	Pflanze	plant	household
schilderij	Gemälde	painting	household
spiegel	Spiegel	mirror	household
stoel	Stuhl	chair	household
stofzuiger	Staubsauger	hoover	household
tafel	Tisch	table	household
waterkoker	Wasserkocher	kettle	household
aansteker	Feuerzeug	lighter	tools
auto	Auto	car	tools
batterij	Batterie	battery	tools
bus	Bus	bus	tools
cd	CD	CD	tools
fiets	Fahrrad	bicycle	tools
helm	Helm	helmet	tools
laptop	Laptop	laptop	tools
lepel	Löffel	spoon	tools
mes	Messer	knife	tools
microfoon	Mikrophon	microphone	tools
mobiel	Handy	mobile	tools
muis	Maus	mouse	tools
pen	Kugelschreiber	pen	tools
printer	Drucker	printer	tools
radio	Radio	radio	tools
schaar	Schere	scissors	tools
sleutel	Schlüssel	key	tools
telefoon	Telefon	telephone	tools
televisie	Fernsehen	television	tools
toilet	Toilette	toilet	tools
trein	Zug	train	tools
vliegtuig	Flugzeug	airplane	tools
vork	Gabel	fork	tools
wasmachine	Waschmaschine	washing machine	tools

APPENDIX B: STIMULI FOR THE PHONOLOGICAL WORKING MEMORY TASK

Practice items:

- toes
- juufoot
- jeemeuboovaus

Test items with three syllables:

- joekeewaup
- waafienoech
- ruufaumiek
- doolieneuf

Test items with four syllables:

- beepoetaamuuf
- hiejeemuutaup
- puudoojienauch
- toopeuriewoem

Test items with five syllables:

- hiepeuloefuuteem
- baawookuujiezaun
- wuutiemoobeejoon
- fooneuwuuzoetaam

Test items with six syllables:

- kootaafieluuzeupiem
- feupaaniezuubeewoes
- waaduukeenoeleumaap
- jienoerooheuwuumaun

APPENDIX C: DESCRIPTIVE STATISTICS SPLIT BY ALL VARIABLE LEVELS

Table C shows means, standard deviations and 95% CIs for all subcombinations of the levels of the four variables included in this study. In this table, the header *Testing moment* covers both the independent variables Exposure frequency and Retention interval.

Table C. Percentage of correctly produced phonemes per target word across all variable levels.

Group	Cognate status	Testing moment	Lag	*n*	Mean	*SD*	95% CI
Experimental	Cognate	EF2	3 trials	30	67.55	20.06	60.06 – 75.04
			7 trials	30	69.71	17.74	63.08 – 76.34
		EF4	3 trials	30	79.16	20.76	71.40 – 86.91
			7 trials	30	84.64	13.08	79.75 – 89.53
		20 minutes	3 trials	30	69.64	22.17	61.36 – 77.91
			7 trials	30	70.79	22.19	62.50 – 79.07
		6 months	3 trials	18	35.03	24.04	23.08 – 46.99
			7 trials	18	36.90	21.62	26.14 – 47.65
	Non-cognate	EF2	3 trials	30	51.04	26.99	40.96 – 61.12
			7 trials	30	41.42	23.37	32.70 – 50.15
		EF4	3 trials	30	72.17	22.66	63.71 – 80.63
			7 trials	30	55.11	26.21	45.32 – 64.90
		20 minutes	3 trials	30	40.46	25.73	30.85 – 50.07
			7 trials	30	40.56	18.56	33.63 – 47.49
		6 months	3 trials	18	9.62	12.74	3.29 – 15.95
			7 trials	18	12.17	19.10	2.67 – 21.66
Control	Cognate	EF2	3 trials	15	12.32	17.34	2.72 – 21.93
			7 trials	15	5.00	7.64	0.77 – 9.23
		EF4	3 trials	15	11.18	12.63	4.18 – 18.17
			7 trials	15	5.40	10.01	-0.15 – 10.94
		20 minutes	3 trials	15	10.00	12.17	3.27 – 16.74
			7 trials	15	6.08	9.50	0.81 – 11.34
	Non-cognate	EF2	3 trials	15	1.89	4.67	-0.70 – 4.47
			7 trials	15	1.11	4.30	-1.27 – 3.49
		EF4	3 trials	15	1.89	4.67	-0.70 – 4.47
			7 trials	15	1.11	4.30	-1.27 – 3.49
		20 minutes	3 trials	15	1.89	4.67	-0.70 – 4.47
			7 trials	15	1.11	4.30	-1.27 – 3.49

Note. EF2 and EF4 refer to the tests that took place during the price comparison task after two and four exposures respectively. 20 minutes and 6 months are the retention intervals for the two post-tests.

APPENDIX D: MODEL COMPARISONS

Tables D and E contain the model comparisons we performed for finding the best random-effects structures for our learning model (Table D) and retention model (Table E).

Table D. Comparing models with different random slopes for modelling the learning data.

Random slope	Converged?	AIC	Test statistics	All dimensions supported?	Included in model?
None	Yes	6999.5	N/A	Yes	N/A
G | W	Yes	6579.1	$\chi^2 = 424.37$, $df = 2$, p **< .001**	Yes	Yes
CS | P	Yes	6427.6	$\chi^2 = 155.52$, $df = 2$, p **< .001**	Yes	Yes
EF | W	Yes	6363.8	$\chi^2 = 69.73$, $df = 3$, p **< .001**	Yes	Yes
EF | P	Yes	6308.6	$\chi^2 = 61.26$, $df = 3$, p **< .001**	Yes	Yes
L | W	Yes	6015.2	$\chi^2 = 301.43$, $df = 4$, p **< .001**	Yes	Yes
L | P	Yes	5751.2	$\chi^2 = 271.95$, $df = 4$, p **< .001**	Yes	Yes
G : EF | W	Yes	5751.2	$\chi^2 = 10.01$, $df = 5$, $p = .07$	Yes	No
CS : EF | P	Yes	5742.5	$\chi^2 = 18.66$, $df = 5$, p **= .002**	Yes	Yes
G : L | W	Yes	5679.3	$\chi^2 = 73.22$, $df = 5$, p **< .001**	Yes	Yes
CS : L | P	Yes	5429.5	$\chi^2 = 261.84$, $df = 6$, p **< .001**	Yes	Yes
EF : L | W	Yes	5375.8	$\chi^2 = 65.68$, $df = 6$, p **< .001**	Yes	Yes
EF : L | P	Yes	5355.6	$\chi^2 = 34.17$, $df = 7$, p **< .001**	Yes	Yes
G : EF : L | P	Yes	5355.0	$\chi^2 = 16.65$, $df = 8$, p **= .03**	Yes	Yes

Note. CS = Cognate status, EF = Exposure frequency, G = Group, P = Participant, L = Lag, W = Word. All models had the same fixed-effects structure, and random intercepts for participants and words: (Number of correct phonemes, Number of incorrect phonemes) ~ G + CS + EF + L + G:CS + G:EF + G:L + (1 | P) + (1 | W). In the *Test statistics* column, each model is compared to the model in the row directly above, provided all criteria for including that above random slope in the model were met. Significant *p*-values are printed in bold.

Table E. Comparing models with different random slopes for modelling the retention data.

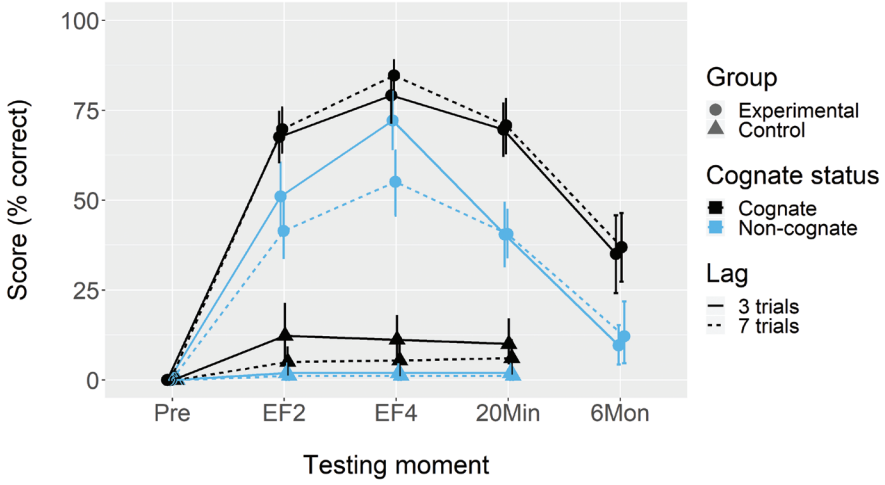
Random slope	Converged?	AIC	Test statistics	All dimensions supported?	Included in model?
None	Yes	7687.6	N/A	Yes	N/A
CS | P	Yes	7566.2	$\chi^2 = 125.41$, $df = 2$, $p < .001$	Yes	Yes
RI | W	Yes	7088.2	$\chi^2 = 487.99$, $df = 5$, $p < .001$	Yes	Yes
RI | P	Yes	6891.4	$\chi^2 = 210.82$, $df = 7$, $p < .001$	Yes	Yes
L | W	Yes	6668.6	$\chi^2 = 230.74$, $df = 4$, $p < .001$	Yes	Yes
L | P	Yes	6571.1	$\chi^2 = 107.58$, $df = 5$, $p < .001$	Yes	Yes
CS : RI | P	Yes	6502.0	$\chi^2 = 95.10$, $df = 13$, $p < .001$	Yes	Yes
CS : L | P	Yes	6266.9	$\chi^2 = 251.00$, $df = 8$, $p < .001$	No	No
RI : L | W	Yes	6326.8	$\chi^2 = 197.13$, $df = 11$, $p < .001$	Yes	Yes
RI : L | P	No	N/A	N/A	N/A	No

Note. CS = Cognate status, G = Group, L = Lag, P = Participant, RI = Retention interval, W = Word. All models had the same fixed-effects structure, and random intercepts over participants and words: (Number of correct phonemes, Number of incorrect phonemes) ~ CS + RI + L + (1 | P) + (1 | W). In the *Test statistics* column, each model is compared to the model in the row directly above, provided all criteria for including that above random slope in the model were met. Significant *p*-values are printed in bold.

APPENDIX E: EXPLORATIVE MODELS

In this appendix, we ran additional models to explore all possible interaction effects, of whom at least some seemed to be present in a visual inspection of Figure 1 from the Results section. For readability, Figure 1 is reprinted here.

Figure 1. Mean scores across four testing moments (EF = Exposure frequency). Error bars represent 95% confidence intervals based on a bootstrap.



Visual inspection of Figure 1 suggests that there was an interaction between Group, Cognate status, Exposure frequency and Lag during the learning phase: In the experimental group the scores for cognate words were higher when learned with a lag of seven rather than three trials, but for non-cognate words it was the other way around. The effect of Lag also seemed stronger after four exposures as compared to after two. To investigate these potential interactions, the explorative learning model had a fixed-effects structure that included all possible interactions between the fixed effects. The random-effects structure was identical to that of the hypothesis-based model as reported in the main text (section 3.3.3). The maximum number of iterations was set to 1,000,000 (because the number of iterations is a function of the numbers of parameters, and should not be less than ten times the number of parameters squared; Bates, Mächler, Bolker & Walker, 2015). The results are shown in Table F.

Table F. Outcomes of the explorative learning model.

Fixed effects		**Logit**	**Odds ratio**	***SE***	***z***	***p***
(Intercept)		2.36	10.64	0.76	3.11	**.002**
G = Control		-11.07	< .001	2.13	-5.19	**< .001**
CS = Non-cognate		-2.29	0.10	0.95	-2.40	**.02**
EF = 4 times		2.19	8.90	0.60	3.62	**< .001**
L = 7 trials		0.01	1.01	0.76	0.02	.99
G = Control : CS = Non-cognate		-3.15	0.04	2.59	-1.22	.22
G = Control : EF = 4 times		-3.03	0.05	1.05	-2.88	**.004**
G = Control : L = 7 trials		-4.82	0.008	3.32	-1.45	.15
CS = Non-cognate : EF = 4 times		0.66	1.93	0.78	0.84	.40
CS = Non-cognate : L = 7 trials		-0.82	0.44	1.11	-0.74	.46
EF = 4 times : L = 7 trials		-0.78	0.46	0.82	-0.95	.34
G = Control : CS = Non-cognate : EF = 4 times		-1.35	0.26	1.83	-0.73	.46
G = Control : CS = Non-cognate : L = 7 trials		1.00	2.71	5.32	0.19	.85
G = Control : EF = 4 times : L = 7 trials		1.75	5.74	1.62	1.08	.28
CS = Non-cognate : EF = 4 times : L = 7 trials		-0.75	0.47	1.00	-0.75	.45
G = Control : CS = Non-cognate : EF = 4 times : L = 7 trials		2.75	15.66	3.18	0.87	.39
Random effects		**Vari-ance**	***SD***			
Participant	(Intercept)	5.12	2.26			
	CS = Non-cognate	4.39	2.10			
	EF = 4 times	1.79	1.34			
	L = 7 trials	5.43	2.33			
	CS = Non-cognate : EF = 4 times	3.55	1.89			
	CS = Non-cognate : L = 7 trials	15.79	3.97			

	EF = 4 times : L = 7 trials	5.90	2.43
	CS = Non-cognate : EF = 4 times : L = 7 trials	6.95	2.64
Word	(Intercept)	10.46	3.24
	G = Control	45.79	6.77
	EF = 4 times	4.19	2.05
	L = 7 trials	7.63	2.76
	G = Control : L = 7 trials	121.04	11.00
	EF = 4 times : L = 7 trials	6.49	2.55

Note. The intercept represents the following combination of predictor 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.

The first thing to note is that those effects that were already estimated with the hypothesis-based model seem quite robust. In other words, the explorative model that included all possible interactions for the most part yielded similar logit estimates, and the significance of the effects was the same across the two models. None of the additional interactions reached significance.

We also ran an explorative model containing all possible fixed-effects interactions for the retention phase. Like the hypothesis-based retention model, this model was computed with the data of the experimental group only. The outcomes are given in Table G.

Table G. Outcomes of the explorative retention model.

Fixed effects		**Logit**	**Odds ratio**	***SE***	***z***	***p***
(Intercept)		4.04	56.64	0.80	5.04	**< .001**
CS = Non-cognate		-1.55	0.21	1.00	-1.55	.12
RI = 20 minutes		-2.14	0.12	0.51	-4.19	**< .001**
RI = 6 months		-5.83	0.003	1.15	-5.07	**< .001**
L = 7 trials		-0.41	0.67	0.62	-0.66	.51
CS = Non-cognate : RI = 20 minutes		-0.60	0.55	0.58	-1.04	0.30
CS = Non-cognate : RI = 6 months		-4.90	0.007	2.13	-2.30	**.02**
CS = Non-cognate : L = 7 trials		-1.26	0.28	0.69	-1.82	.07
RI = 20 minutes : L = 7 trials		0.40	1.50	0.50	0.81	.42
RI = 6 months : L = 7 trials		0.53	1.69	1.16	0.45	.65
CS = Non-cognate : RI = 20 minutes : L = 7 trials		1.34	3.81	0.56	2.40	**.02**
CS = Non-cognate : RI = 6 months : L = 7 trials		1.65	5.20	1.68	0.98	.33
Random effects		**Vari-ance**	***SD***			
Participant	(Intercept)	2.37	1.54			
	CS = Non-cognate	1.75	1.32			
	RI = 20 minutes	0.66	0.81			
	RI = 6 months	5.20	2.28			
	L = 7 trials	0.90	0.95			
	CS = Non-cognate : RI = 20 minutes	0.70	0.84			
	CS = Non-cognate : RI = 6 months	28.53	5.34			
Word	(Intercept)	13.40	3.66			
	RI = 20 minutes	2.68	1.64			
	RI = 6 months	23.81	4.88			
	L = 7 trials	4.80	2.19			
	RI = 20 minutes : L = 7 trials	1.88	1.37			
	RI = 6 months : L = 7 trials	24.45	4.95			

Note. The intercept represents the following combination of predictor levels: CS [Cognate status] = Cognate, RI [Retention interval] = 4 exposures (i.e., participants' scores after 20 minutes and 6 months are compared to their last score from the learning phase), 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.

In this case, there is an obvious difference to the hypothesis-based retention model that contained only main effects in the fixed-effects structure: The cognate effect is not significant in the explorative model. However, upon closer inspection this seemingly surprising finding is easy to explain. Because the explorative model included all possible interactions between Cognate status and the other predictors, the main effect of Cognate status (with p = .12)

was only computed for EF4 at Lag 3 (in the hypothesis-based model, it was computed by averaging over both exposure frequencies and lags). When we look at Figure A, it can be seen that the experimental participants' scores for cognates and non-cognates, at EF4 with Lag 3, are close together and have overlapping confidence intervals. However, and in line with Figure A, the output in Table G indicates that the cognate effect was significantly different after six months, as well as after 20 minutes when considering Lag 7.

APPENDIX F: INTERPRETATION OF LOGIT AND ODDS RATIO

The logit estimates in Tables 4 and 5 in the main text are approximations of the probability that a phoneme in a target word is produced correctly in various conditions (e.g., in cognate versus non-cognate words). The model estimates themselves are not expressed as probabilities, since probabilities always lie between 0 and 1 while the predictions of linear models are not limited to this range. Rather, the estimates are expressed as the logarithm of the odds ('logit') by $x = \log(p / (1 - p))$, and can be transformed back to probabilities through the formula $p = e^x / (1 + e^x)$, where x is the logit and e is a mathematical constant, approximately equal to 2.72. Say we wanted to know how the learning model estimates the probability that participants in the experimental group correctly produce a phoneme in a non-cognate word they were exposed to four times with a lag of three trials. According to the model's predictions reported in Table 4, the logit would be 2.80 – 3.25 + 1.72 = 1.27, and the corresponding estimated probability would therefore be $2.72^{1.27} / (1 + 2.72^{1.27}) = 0.78$.

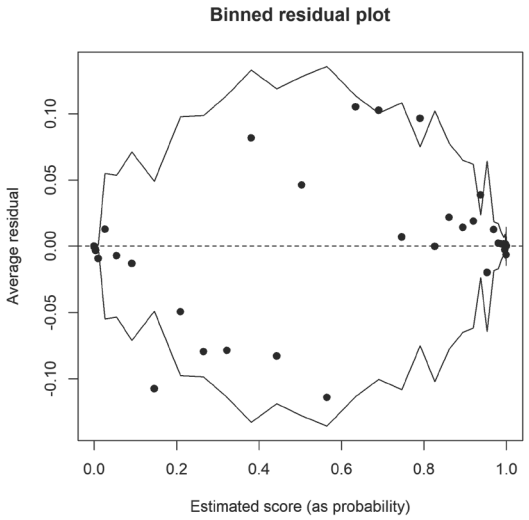
Effect sizes are expressed as odds ratios (ORs). With the exception of the intercept itself, the OR tells us how the odds of correctly producing a phoneme change for one predictor level as compared to the level of that predictor that is represented by the intercept. For example, in Table 4, the odds that a phoneme is produced correctly after four exposures are 5.60 times higher than after two exposures. As far as we know, there are no guidelines yet for interpreting OR magnitudes in L2 research. However, an online search for guidelines in other fields revealed that generally speaking, ORs around 1.5 are interpreted as small, ORs between 2.5-3.5 as medium, and ORs bigger than 4-9 as large (see Footnote 4 in Chapter 4).

ORs under 1 should first be converted before applying the above guidelines. For example, Table 4 shows that the odds to correctly produce a phoneme in a non-cognate word are 0.04 times higher than the odds to correctly produce a phoneme in a cognate word. Thus, this means that the former odds are in fact much smaller than the latter odds. We can turn the tables by dividing 1 by the OR: The odds to correctly produce a phoneme in a cognate word are 25 (i.e., 1/0.04) times higher than the odds to correctly produce a phoneme in a non-cognate word. Thus, ORs of 0.04 and 25 are equivalent, but represent different directions of a certain effect. Overall, the further the OR is removed from 1 (in either direction), the stronger is the effect.

APPENDIX G: EVALUATING MODEL FIT

Evaluating model fit for logistic models is not straightforward. Plotting raw residuals, as one might do for data with continuous outcomes, is not very informative in this case, as a $residual_i$ can only take one of two values, depending on y_i (namely 0 or 1). A solution is to create bins of residuals based on their fitted values, and plot the average of each bin (Gelman & Hill, 2007, p. 97). This was done for the learning model in Figure A, which was created with the *arm* package in R (version 1.9-3; Gelman et al., 2016).

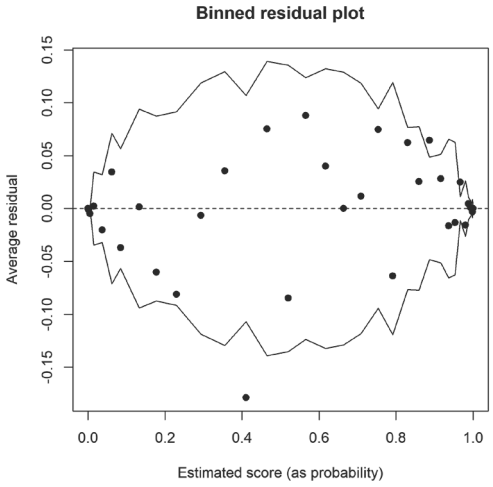
Figure A. Binned residual plot for the learning model.



The x-axis represents the scores estimated by the learning model. There are 45 bins (taking the square root of the number of data points (2022) is the default in R). ±2 standard-error bounds were computed as $2\sqrt{p(1-p)/n}$ (Gelman & Hill, 2007, p. 97); they would be expected to contain 95% of binned residuals, which indeed seems to be the case. However, it can also be seen that the residuals are not uniformly distributed, as they should be (on average, they have negative values for lower scores, and positive values for larger scores).

Still, as compared to models with different random-effects structures, our learning model achieved a better fit to the data in terms of log-likelihood. For instance, we ran another hypothesis-based model with a simpler random-effects structure containing only main effects, but no interactions: (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 + Exposure frequency + Lag | Word). Its binned residual plot looks a little better (Figure B):

Figure B. Binned residual plot for another learning model with a simpler random-effects structure.



However, our original model fitted the data significantly better ($\chi^2 = 470.22$, $df = 37$, $p < .001$). The AIC score of the original model was 5355.0 and of the simpler model it was 5751.2 (lower AIC scores represent better models). Thus, we conclude that even though the errors were not uniformly distributed in our original model, it still provided a better fit to our data than a model with a more uniform distribution of errors, and therefore we have reported the results for this best-fitting model in Table 4 in the Results section. Reassuringly, the logit estimates and significance values were similar for both of these learning models, and the conclusions drawn from both models would be the same.

The binned residual plot for the retention model is shown in Figure C.

Figure C. Binned residual plot for the retention model.

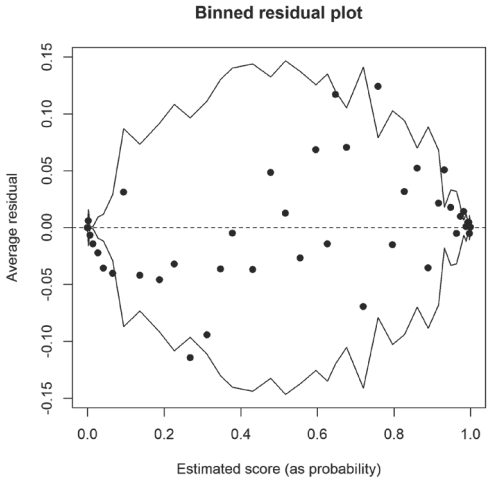


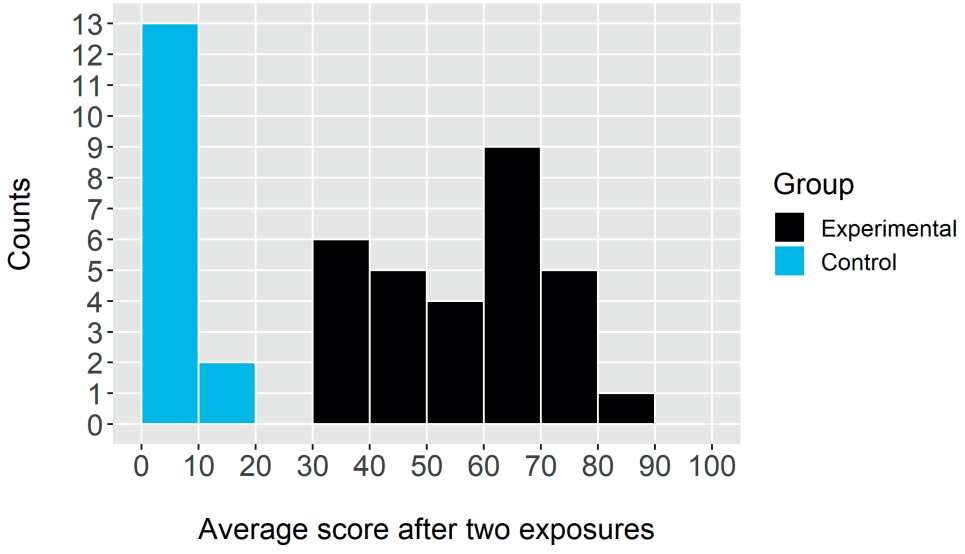
Figure C shows that the distribution of the residuals is relatively uniform, which is good. However, the figure also shows that the model struggled when predicting values under ±0.1, which it systematically overestimated (to a lesser extent, this could also be seen for the learning model). Upon further inspection, this is not surprising. Our data set did not contain any observations of scores under 0.14 (except the score of 0, which was assigned when participants could not produce a word). Such a low score, for example 0.10, would mean that a participant produced one out of 10 phonemes correctly. Most of our target words were not that long, and it would be unusual anyway to have so little knowledge of a word and still be able to produce something at all.

Since our model made continuous predictions, it seems that it sometimes predicted some knowledge of words which the participants in reality had zero knowledge of. This pattern in the residuals was also visible in retention models with simpler random-effects structures. Thus, no better alternative was available. However, the situation does not seem to be very problematic, since we were focused on investigating contrasts rather than the absolute values of predictions.

APPENDIX H: ANALYSIS OF INDIVIDUAL DIFFERENCES

In this appendix, we will look into individual differences between participants. Figure D shows a histogram of the scores, averaged over words, that were obtained after two exposures in the price comparison task. Most participants in the control group either scored at 0 or close to 0. The scores in the experimental group are relatively normally distributed, with no real outliers.

Figure D. Histogram of average participant scores obtained after two exposures, split by Group.



As described in the Methods section (3.2.2.2), we tested the participants on a variety of measures to identify individual differences. We can use these outcomes to gain more insight in the relationship between individual characteristics and L2 word learning abilities. In this analysis, we will use the data from the experimental group only, because the control group did not have actual opportunities to learn words. Table H shows the correlations between the measures of individual differences and the learning scores after two exposures. The correlations between the measures of individual differences are also shown. We used Pearson's correlation coefficient (r_p) if the data of both measurements were normally distributed (as shown by a Shapiro-Wilk test), and Spearman's rho (r_s) otherwise.

Table H. Correlations between measurements (experimental group only).

	Score (EF2)	Age	Years of learning Dutch	Self-rated proficiency	Amount of daily exposure to Dutch	Number of other languages known	Dutch vocabulary (LexTALE)	Phonological working memory
Score (EF2)	-	-	-	-	-	-	-	-
Age	r_s = .52 **p < .001**	-	-	-	-	-	-	-
Years of learning Dutch	r_s = .43 **p < .001**	r_s = .68 **p < .001**	-	-	-	-	-	-
Self-rated proficiency	r_s = .50 **p < .01**	r_s = .44 **p = .02**	r_s = .37 **p = .04**	-	-	-	-	-
Amount of daily exposure to Dutch	r_s = .26 p = .16	r_s = .05 p = .78	r_s = .06 p = .74	r_s = .36 p = .054	-	-	-	-
Number of other languages known	r_s = .20 p = .29	r_s = .00 p = .99	r_s = -.13 p = .48	r_s = -.02 p = .91	r_s = -.01 p = .94	-	-	-
Dutch vocabulary (LexTALE)	r_s = .34 p = .06	r_s = .39 **p = .01**	r_s = .14 p = .46	r_s = .26 p = .17	r_s = .12 p = .53	-	-	-
Phonological working memory	r_p = -.23 p = .23	r_s = -.35 p = .06	r_s = -.33 p = .07	r_s = -.13 p = .50	r_s = -.16 p = .40	r_s = -.17 p = .37	r_p = .09 p = .64	-

Note. N = 45, except for the correlations involving phonological working memory, where N = 44 because one participant did not complete the phonological working memory test. p-values under .05 are printed in bold (one might still want to apply a correction for multiple testing).

As can be seen, three predictors significantly predicted L2 word learning abilities: Age, Years of learning Dutch, and Self-rated proficiency. These findings are in line with the so-called Matthew effect ("the rich get richer"). In the current context, this means that more proficient learners have an advantage when it comes to L2 word learning. Such advantages have often been found in the literature (e.g., Montero Perez, Peters & Desmet, 2014; Vidal, 2011; Vulchanova, Aurstad, Kvitnes & Eshuis, 2015). In our own data, we also see that Dutch vocabulary size was weakly related to word learning with r_s = .34, although this predictor did not reach significance.

The predictors of Age, Years of learning Dutch, and Self-rated proficiency are all correlated among themselves as well, the highest correlation being r_s = .68 between Age and Years of learning Dutch. This correlation likely is explained by the fact that most of the participants started to learn Dutch around the age of 19, after having finished high school and having moved to the Netherlands to study at university. In general, learners who have spent more years learning Dutch on average will be older and more proficient in Dutch.

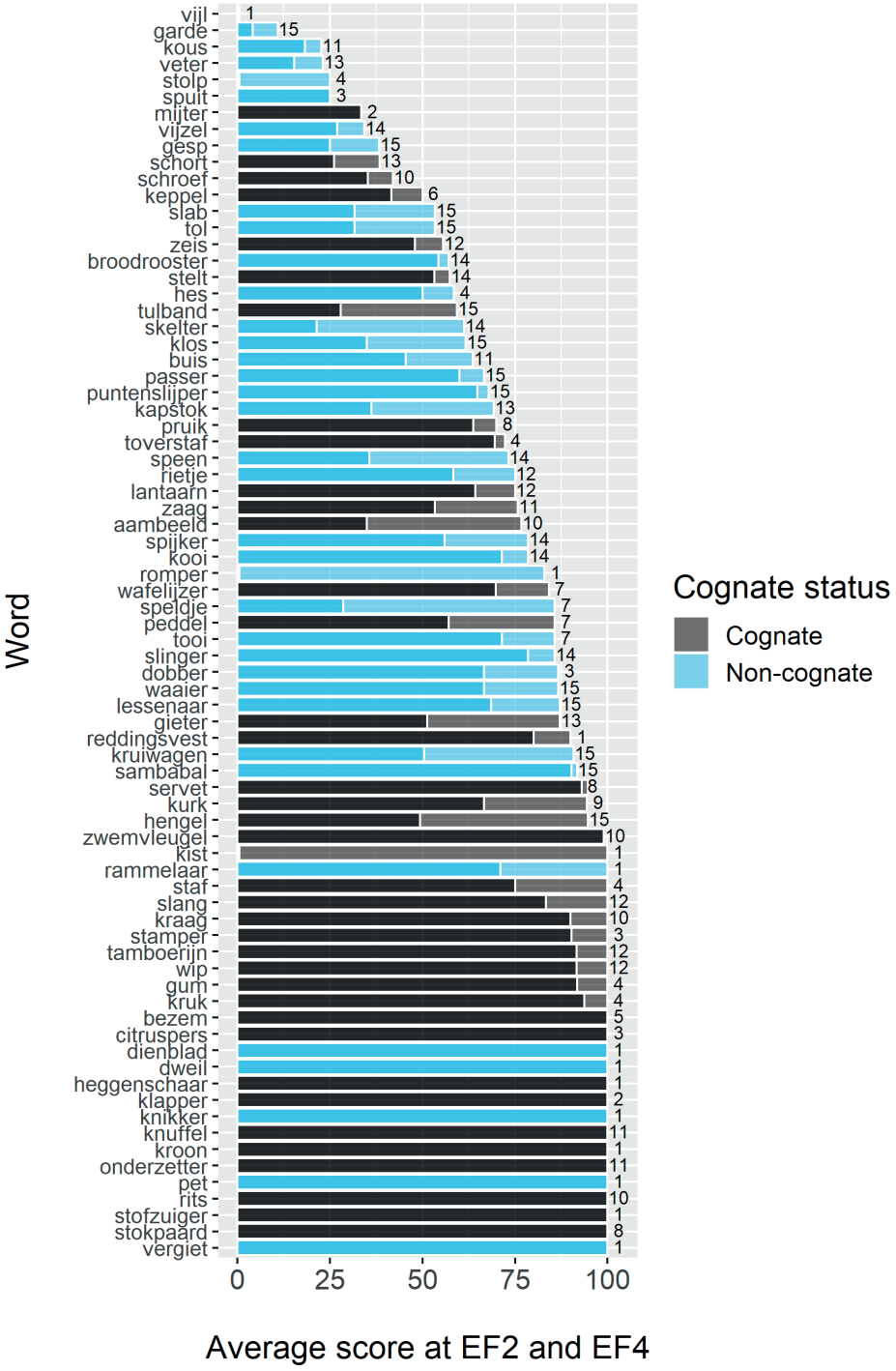
APPENDIX I: ANALYSIS OF TARGET ITEMS

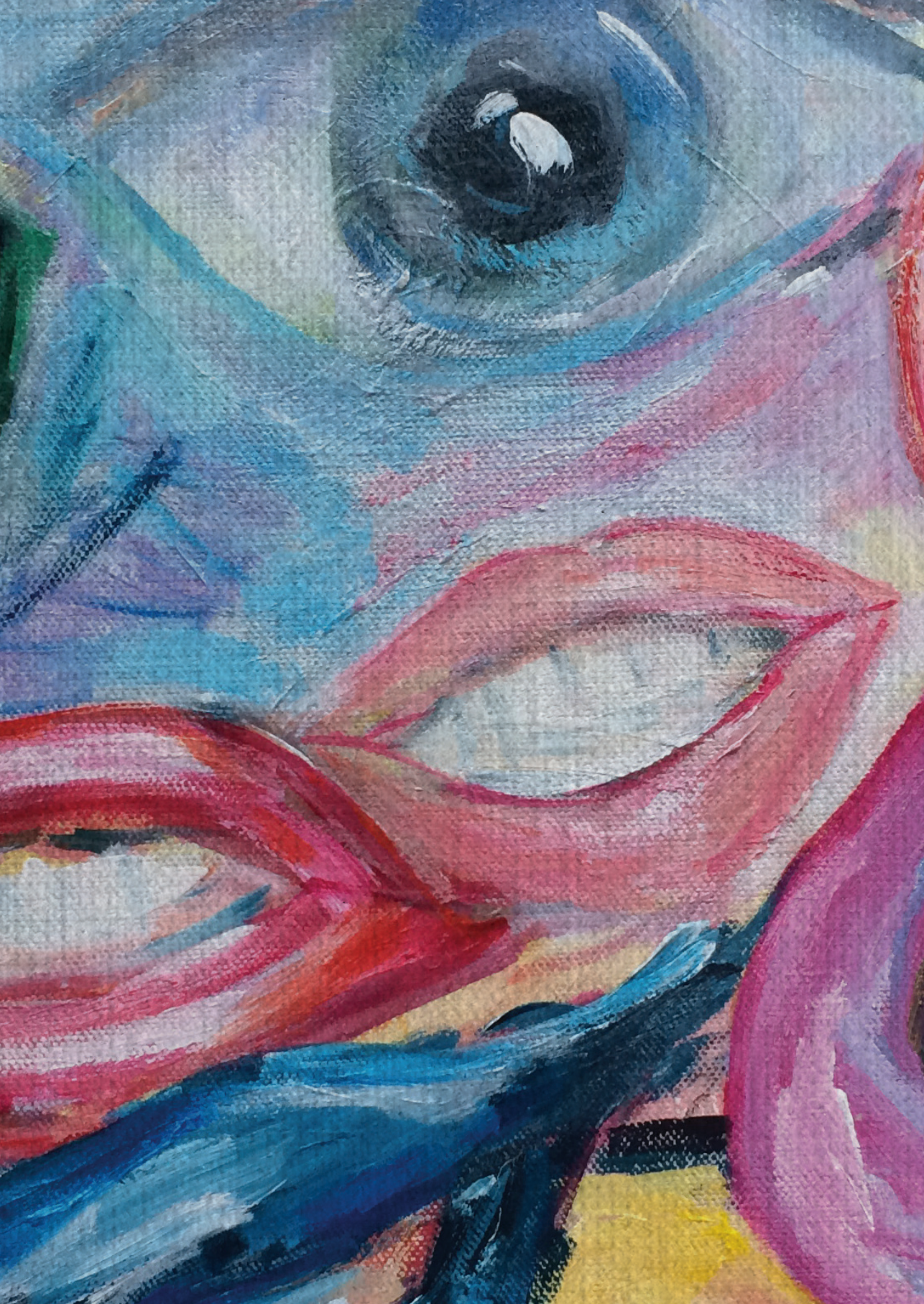
Figure E below shows the average scores per word in the experimental group. The darker part of each bar represents the average score over participants after two exposures, and the lighter part of each bar represents the participants' scores after four exposures. The number to the right of each bar indicates for how many participants the particular word was included in their set of unknown target items. Since there were 30 participants in the experimental group, and half of them learned cognates in a given semantic category and the other half non-cognates, the maximum possible *n* in Figure E is 15.

As an example, the second word in the graph, *garde* (English: *whisk*), was part of the target item set of 15 participants. This means that none of the 15 participants who were pre-tested on this item knew it productively in advance of the experiment. For all these participants, *garde* was selected as one of the six to-be-learned words in the household semantic category. In contrast, the first word, *vijl* (English: *file*), was in the target item set for only one participant. This means that this particular participant already knew one or more of the six default words in the tools semantic category in the pre-test, and this word was replaced by *vijl*, which the participant did not know in the pre-test.

The graph is sorted by the average score obtained at after four exposures. It can be seen that at this point, about 30% of the words (from *kist* (English: *chest*) onwards) were learned perfectly by all of the participants for whom this word was in their target item set. Apparently, these items were relatively easy. The words from *bezem* (English: *broom*) onwards even had been learned perfectly by all participants already after two exposures. There are relatively many cognates among these words. All the non-cognates for which a perfect score was achieved after two exposures were only in the target item set of one participant. Therefore, these estimations are more uncertain and could potentially be flukes.

Figure E. Scores per word (target items only) that were obtained after two and four exposures, averaged over the participants in the experimental group. Dark bars represent scores after two exposures, and light bars represent scores after four exposures.





4.

Noticing vocabulary holes aids incidental second language word learning: An experimental study

This chapter is based on:

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

ABSTRACT

Noticing the hole (NTH) occurs when speakers want to say something, but realise they do not know the right word(s). Such awareness of lacking knowledge supposedly facilitates the acquisition of the unknown word(s) from later input (Swain, 1993). We tested this claim by experimentally inducing NTH in a second language (L2) for some participants (experimental), but not others (control). Then, in a price comparison task, all participants were exposed to spoken L2 input containing the to-be-learned words. They were unaware of taking part in an L2 study. Post-tests showed that participants who had noticed holes in their vocabulary had indeed learned more words as compared to participants who had not. This held both for the experimental group as well as for those participants in the control group who later reported to have noticed holes. Thus, when we become aware of vocabulary holes, the first step to improve our vocabulary is already taken.

4.1 INTRODUCTION

Second language (L2) learners often fail to exactly express their intended message, due to a lack of knowledge of the target language vocabulary. This is especially poignant in real-life conversations, where there is little occasion to consult a dictionary whilst speaking. Although learners are usually able to talk around their lacking word knowledge, the forced resort to circumlocution may not go unnoticed by the learners themselves.

While the awareness of being at a loss for words may be frustrating, it may well be beneficial to the second language acquisition (SLA) process. This possibility underlies one of the four hypothesised functions of output, according to Swain's Output Hypothesis (1985, 1993, 1995, 1998), namely its *noticing function* (the other functions would be practicing, hypothesis testing, and the metalinguistic function). When learners fail to produce target language output, be it vocally or subvocally (Swain, 1995, p. 125), this "may prompt [them] to consciously recognise some of their linguistic problems" (Swain & Lapkin, 1995, p. 373). In turn, this could trigger cognitive processes involved in SLA, such as a heightened state of attention for subsequent input (Swain & Lapkin, 1995, p. 386), which may be beneficial to learning.

Swain's use of the term *noticing* differs from how it was originally used by Schmidt (1990) in his Noticing Hypothesis, which states that noticing would be a necessary condition for language learning. Schmidt (2001, p. 4) equates noticing to "awareness at a very low level of abstraction": learners' awareness of specific instances in the language input. For example, learners may notice how native speakers use a particular form in the target language (Izumi, 2013, p. 38). If learners also compare their own imperfect use of that form to the way the more proficient speaker used the form in the input, this is called *noticing the gap* (Schmidt & Frota, 1986). We will use the term *noticing (the gap)* to catch both of Schmidt's constructs in one phrase.

While noticing (the gap) concerns learners interacting with external language input, Swain's noticing function of output comes into play when learners struggle to produce language, regardless of whether the output is vocalised or not. This applies to both grammatical structures and words. In this study, we will focus on the latter. When learners become aware that an L2 target word is completely absent in their vocabulary, this is called *noticing the hole in one's interlanguage* (e.g., Doughty and Williams, 1998, p. 255). When learners struggle to produce a word they have incomplete knowledge of, it is called 'noticing the gap in one's ability' (Izumi, 2013, p. 40).

Importantly, *noticing the gap in one's ability* is not the same as Schmidt and Frota's (1986) *noticing the gap*, because the former happens learner-internally and the latter in relation to external input. To avoid confusion in terminology, in this chapter we will speak of *noticing the hole* (NTH) when referring to situations where learners struggle to produce output and become aware of their linguistic problem, be it because a word is completely absent in their vocabulary (a hole in one's interlanguage), or because it is only partially represented (a gap in one's ability).

How can the hypothesised facilitative effects of NTH on vocabulary learning be explained in terms of cognitive mechanisms? Imagine a learner making an unsuccessful attempt to produce a word, thereby experiencing NTH. Suppose that this learner is subsequently exposed to this word. It is hypothesised that the learner will remember the word more readily, as compared to a situation in which he/she did not notice the hole before being exposed to input. This would be an instance of the pre-testing effect observed in memory experiments, where an unsuccessful retrieval attempt before exposure to the relevant materials enhances learning (Grimaldi & Karpicke, 2012; Kornell, Jensen Hays & Bjork, 2009; Richland, Kornell & Kao, 2009).

Several explanations for the pre-testing effect have been proposed, including the impact of unsuccessful retrieval on intentional learning behaviour (Richland et al., 2009): It could well be that failure to produce a word alters intentional learning behaviour by fostering epistemic curiosity, i.e., "the desire for knowledge that motivates individuals to […] eliminate information-gaps" (Litman, 2008, p. 1586). In turn, humans are better at learning information they are curious about (Gruber, Gelman & Ranganath, 2014; also see Kang et al., 2009). Gruber et al. (2014) name attentional processes as one potential explanation of the relationship between curiosity and learning (although they also mention it is likely there are other variables too). For three retrieval-based explanations of the pre-testing effect, see Kornell et al. (2009).

4.1.1 Literature review: From NTH to SLA

In the present study, we experimentally manipulated NTH by confronting German learners of Dutch with their lacking L2 vocabulary knowledge. The study will be introduced in more detail in the next section (4.1.2). Before doing so, we present a literature review of other experimental studies concerning NTH and SLA (for two observational studies, see Hanaoka, 2007, and Hanaoka & Izumi, 2012).

4.1.1.1 Grammar studies

Izumi, Bigelow, Fujiwara and Fearnow (1999) and Izumi and Bigelow (2000) studied the acquisition of the English past hypothetical conditional (e.g., "If Ann had traveled to Spain in '92, she would have seen the Olympics", Izumi et al., 1999, p. 426). Specifically, they investigated whether the anticipation of an output task (here: knowing that one later has to do a writing task), and the actual execution of this output task, lead to noticing and improved acquisition of the target structure.

Izumi et al. (1999, p. 423) indicate that what they call noticing actually encompasses two separate processes: noticing "problems in one's interlanguage" (what we call *NTH*), and noticing "the relevant features in the input" (what we call *noticing (the gap)*). This 'noticing' was measured by letting the participants read a text containing the target structure, and asking them to underline the parts they thought were relevant to their upcoming activity. Only for the experimental group, the upcoming activity was an output task. The control group

knew they would have to answer comprehension questions about the text. After completing their respective activities, the participants did the underlining task again.

In neither experiment did the groups differ significantly in their underlining behaviour. Thus, neither the anticipation of an output task, nor the (presumed) experience of NTH during such an output task, resulted in the learners noticing (the gap to) the target structure more often. Regarding the acquisition of the target structure, the experimental group did significantly outperform the control group in one contrast (out of many) in the 1999 study, with a large effect size of $d = 1.36$.[1] However, one should note that these studies seem to be at risk of both Type-I and Type-II errors, because no correction for multiple testing was applied, and overall sample sizes were rather small ($N = 22$ in 1999, and $N = 18$ in 2000).

A very similar study was conducted by Song and Suh (2008), using the same target structure and tasks. One additional experimental output group was added, which (supposedly) noticed holes through a picture-cued writing task that required use of the target structure. In this study, the participants in the two experimental groups did underline significantly more conditional-related items than the control participants who did not produce written output. It did not matter whether the underlining task took place before or after the output activity. Thus, in this study, anticipating and experiencing NTH in an output task increased the participants' noticing (the gap to) the target structure. It was also shown that scores on a post-test production task were higher in the experimental groups than in the control group ($d = 0.72$ and $d = 0.95$). However, differences on a recognition task were absent. The authors do not address potential reasons for the discrepancies between the outcomes of this study and the earlier studies by Izumi and colleagues.

Two issues relating to the above studies need to be discussed. First, it may be that the activities in the experimental and control groups following exposure to the target structure differed in depth of processing. That is, when writing a text and thereby reproducing the target structure, this structure is likely to be processed more deeply than when answering comprehension questions. The positive relationship between depth of processing and learning has long been posited (Craik & Lockhart, 1972; Craik & Tulving, 1975) and supported, also for SLA (e.g., Laufer & Hulstijn, 2001; Leow, 2015). Therefore, potential differences between the experimental and control groups might to some extent be due to different depths of processing, rather than NTH and noticing (the gap) exclusively.

Differences in processing debt are indeed mentioned in Izumi and Izumi (2004), which is another study that used the above-described design. Unexpectedly, the researchers found that their control group improved more on the target structure than their experimental group. In their discussion, Izumi and Izumi concede that differences in processing depth may have contributed to this unexpected finding. As of yet, however, such alternative explanations cannot be empirically evaluated, because depth of processing was not measured in any

[1] All effect sizes (expressed as Cohen's *d*) mentioned in this Introduction were calculated by the author of this thesis with data from the articles.

of the above studies. Therefore, if researchers choose to use different treatments for the experimental and control groups, they should ideally include measurements of depth of processing to evaluate such alternative explanations.

A second concern is that the adequacy of underlining as a measure of noticing (the gap) is questionable. Song and Suh (2008, p. 308) remark that this method may not be suitable "for tapping into learners' noticing and attention" and that think-aloud or stimulated recall protocols might provide a better solution. Izumi and Bigelow (2000, pp. 270–271) admit that one cannot be sure that underlining captures all items that were attended to, nor that it excludes items that were not attended to. For future studies, they recommend triangulation with other measures.

Such a triangulation was performed by Uggen (2012). Her design was very similar to Izumi and Bigelow (2000), again revolving around the past hypothetical conditional and (the anticipation of) output tasks. This time, there was an additional experimental group, which was trained and tested on the present hypothetical conditional. For the triangulation of noticing measurements, Uggen also analysed the participants' essays qualitatively, and added stimulated recall. Having finished the experimental procedure, her participants watched a video recording of the experimental session and commented on the thoughts they had had at the time. This stimulated recall measurement proved especially valuable, as it showed that in one experimental group the participants also commented on grammatical features that they had not underlined. The underlining measurement itself again was not very useful, as no differences in underlining could be detected between the two experimental groups and the control group. With regard to acquisition, the experimental group that was assigned the past hypothetical conditional showed significant improvement on this structure. The other experimental group, assigned the present hypothetical conditional, did not improve. According to Uggen (2012, p. 533), perhaps this happened because this latter structure was less complex and therefore less "noticeable" to the learners.

In summary, Uggen's (2012) study suggests that written output influences learners' "awareness of their linguistic limitations concerning grammar structures" (p. 506). Considering all studies discussed so far, it seems that NTH can benefit the acquisition of L2 grammatical structures, but that these structures need to be of a certain complexity. Furthermore, to measure noticing (the gap), triangulation of measurements is recommended. Uggen (2012) showed that underlining alone does not suffice.

4.1.1.2 Vocabulary studies

So far we have only discussed studies on L2 grammar learning. The outcomes of these studies may not be directly transferrable to word learning, as grammar learning revolves around learning a rule or pattern, while vocabulary requires memorising word forms. However, the different types of noticing that were discussed above are relevant to both grammar and word learning. After all, both types of learning can be expected to depend on a learner's attention to input (noticing (the gap)), and the learner's awareness of his/her own state of knowledge

(NTH). The current study focuses on NTH. To our knowledge, there are only two studies on the effects of NTH on vocabulary learning, both of which focused on the written domain (Kwon, 2006; Mahmoudabadi, Soleimani, Jafarigohar & Iravani, 2015). Both studies manipulated the order in which participants performed output and input tasks.

In the input task in Mahmoudabadi et al. (2015), the participants connected written words with their corresponding pictures. In the output task, the participants had to name the same pictures, but without a word list. In Kwon (2006), the input and output tasks comprised a variety of activities. The input tasks were reading a text and answering comprehension questions, looking at pictures and answering comprehension questions, and a word recognition task. The output tasks were fill-in-the-blank, answering open questions, and narrative writing. In both studies, it was assumed that the output tasks would elicit NTH (when participants failed to produce the target words). Thus, the participants in the input-before-output conditions were exposed to input containing the target vocabulary *before* having noticed holes, and the participants in the output-before-input conditions *after* having noticed holes.

Vocabulary post-tests were administered after the completion of all output and input tasks. Mahmoudabadi et al. (2015) found a significant facilitative effect (d = 0.98) of NTH, i.e., more word learning in the output-before-input than in the input-before-output condition. Kwon (2006) found no significant effect of NTH. Her preferred explanation (pp. 118–120) for this null result, reminiscent of Doughty's (2001) *cognitive window*, is that the delay between the output and input tasks was too long. This may have weakened any potential effects of NTH.

Leow (1999, p. 66) has pointed out that it cannot automatically be assumed that participants will behave according to the experimental instructions or the experimenter's expectations. Accordingly, in both studies a subsample of the participants was interviewed after the treatment and post-tests. Mahmoudabadi et al. (2015) explicitly asked 10 participants in the output-before-input condition (out of 43) whether they felt the need to know the words when doing the output task. All said yes. Kwon (2006) interviewed a total of 10 participants (sampled from both task orders, out of 80). From the excerpts provided, it seems that at least some of the participants in the output-before-input conditions during the output tasks realised that they did not know the words they needed, and became motivated to find them in the input. It is unclear whether this applies to all participants.

As in the grammar studies, in the vocabulary studies too there seems to be a confound between the NTH manipulation and opportunities for processing the input. The groups did not only differ (as intended) in whether or not the participants were expected to notice holes before exposure to input, but also in their opportunities to process that input. Only the input-before-output group could have benefitted from the retrieval of words from memory during the output task, which has been shown to facilitate vocabulary learning and retention (Barcroft, 2007). The output-before-input groups did not have this opportunity for retrieval

practice, as there was nothing yet to retrieve. This difference between the conditions is conceptually distinct from, but confounded with, the NTH manipulation.

In conclusion, while we do not doubt the relevance of the above-discussed studies for L2 pedagogy, their design makes it difficult to isolate the true effect of NTH on SLA. The present study therefore employed an experimental design in which, after the NTH manipulation, exposure to input and opportunities to process that input were identical in all conditions. Still, keeping Leow (1999) in mind, we realised that we could not assume that NTH happens whenever researchers create a setting where it is expected to occur, and does not happen otherwise. Perhaps this could also explain why some of the above studies did not find significant effects of NTH. To check whether our manipulation worked as expected, we interviewed our participants regarding their NTH experience after the experiment.

4.1.2 The present study: NTH in incidental L2 word learning

We addressed the questions of whether NTH in spoken L2 word production facilitates the acquisition of these words from spoken input, and how well these words are retained over a short period of time. The participants were German native speakers, with Dutch being the L2. We used a task that was advertised as a price judgment task, but unbeknownst to the participants was seeded with low-frequency non-cognate Dutch words. This allowed us to investigate L2 vocabulary learning (more details will follow in the Methods section).

To induce NTH in the experimental condition, we asked the participants to name the objects in Dutch. We expected that the inability to name a given object would result in NTH. Post-experiment interviews showed that this expectation was correct. In contrast, the participants in the control condition inspected the same objects, but did not name them. This ensured that both groups were equally familiar with the materials. The expectation that this silent inspection would not result in NTH was also checked in post-experiment interviews. In fact, it was found that about half of the participants in the control condition had noticed holes after all. Following Leow (2000), we analysed their data separately (see Analysis). To this end, we tested more participants in the control condition, such that we could form separate groups of participants who did and who did not report noticing holes. In this way, we could not only assess the effect of the external, experimental induction of NTH, but also that of the spontaneous internal occurrence of NTH when it was not experimentally induced.

Having named or silently studied the pictures (i.e., after the NTH manipulation), both groups underwent the same procedure. Specifically, the participants were exposed to naturalistic L2 input from a Dutch native speaker in the form of price comparisons. The input contained, in a highly controlled way, the names of the objects previously unknown to the participants. The participants were unaware that they were expected to learn from this input and would later be tested on it.

After the exposure to input, the participants took two unannounced post-tests (immediately and after 15 minutes) to measure how many words they had learned and retained. The 15-minute interval allows us to study the earliest stages of the forgetting

curve, as shown for the first time in a classic experiment by Ebbinghaus (1885/1913/2011). Ebbinghaus memorised lists of nonsense syllables (e.g., *zup*). Having studied the lists until he reached a score of 100% correct, 20 minutes later he could only remember 58%. This shows how rapidly newly-acquired knowledge can decay.

Post-experiment interviews confirmed that the participants were indeed unaware of the study's language learning aspect. Thus, with this task we can approach real-life incidental L2 word learning in the laboratory, while maintaining a high degree of experimental control.

4.2 METHODS

4.2.1 Participants

The participants were 70 German students in Nijmegen, the Netherlands. Crucially, they did not know the study was targeted at German native speakers, as it was advertised as a psychological experiment about making price judgments. Non-German participants were prevented from signing up through a hidden language filter in the online participant recruitment system. Thus, the participants were fully naive regarding the language aspects of the study.

The participants were randomly assigned to the experimental and control condition. In the experimental condition, the participants tried to produce output before being exposed to the target words, and therefore noticed holes (as confirmed through interviews at the end of the experiment). We will call this condition [+O, +NTH] (see Procedure for more details on the manipulation). It should be noted that "+O" (+Output) mainly reflects situations where participants *tried* to vocally produce output, but actually failed to do so. In the control condition, the participants were not required to produce output before the exposure to input, and thus were supposed not to notice holes: [−O, −NTH]. However, the post-experiment interviews revealed that almost half of the participants in the control condition had nevertheless noticed holes, as they had internally tried to name the target items. These participants were assigned to a new, third condition, which was called [−O, +NTH]. Thus, while +/- O was experimentally controlled, +/- NTH resulted from individual differences (in the original control condition only, as everyone in the experimental condition noticed holes). Testing was continued until all three conditions included a minimum of 20 participants whose data could be used.

Four participants were excluded from the analysis because they indicated during the second post-test that they had already actively known more than 25% of the target words before the experiment (see Debriefing and measures). One additional participant was excluded because he had not understood the price judgment task. The final sample thus included 65 participants (51 females), who had all been raised with German as their only native language. The participants' mean age was 22 (range 19–27); they had started learning Dutch at a mean age of 19 (range 16–24). All but one were, at the time of this study, taking higher education courses taught in Dutch, or had done so in the past. In addition to German and Dutch, all participants reported knowledge of English, and some reported knowledge of

additional languages. None of the participants in the final sample guessed the purpose of the study (Dutch word learning) during debriefing.

The participants in the three conditions were compared by means of one-way independent ANOVAs on a number of dimensions that could potentially influence L2 word learning (see Table 1). Prior Dutch vocabulary size was determined with the Dutch version of the LexTALE vocabulary test (**www.lextale.com**; also see Lemhöfer & Broersma, 2012). To get an impression of the participants' motivation and strategy use in learning Dutch, they were asked to rate a number of statements (shown to them in German) on a 1–5 scale. We selected four of these statements for our analysis, namely: 1) "It is important to me to have a large Dutch vocabulary", 2) "The way in which something is said is not important to me, only what it means", 3) "When I hear a Dutch word I do not know, I try to learn it", and 4) "I pay attention to subtle differences between German and Dutch". All variables except LexTALE and Passive knowledge of target words were gathered through a background questionnaire that the participants completed after the experiment (see Debriefing and measures). Table 1 shows no significant differences between the groups in any of the measures (all *p*s > .13).

Table 1. Mean scores and standard deviations (in parentheses) of participant characteristics in the three conditions.

	[+O, +NTH] *n* = 21	[-O, +NTH] *n* = 20	[-O, -NTH] *n* = 24	Test statistics
Age	23.00 (2.21)	22.25 (1.92)	22.25 (2.23)	$F(2,62) = 0.88$, $p = .42$
Years of learning Dutch	3.38 (2.94)	2.66 (1.57)	2.59 (1.64)	$F(2,62) = 0.91$, $p = .41$
Self-rated proficiency*	3.52 (0.60)	3.55 (0.69)	3.42 (0.65)	$F(2,62) = 0.27$, $p = .77$
Current amount of exposure to Dutch*	3.11 (0.46)	3.55 (0.87)	3.38 (0.83)	$F(2,62) = 1.80$, $p = .17$
Number of other languages known	2.38 (0.59)	2.35 (0.75)	2.33 (0.76)	$F(2,62) = 0.03$, $p = .97$
Statement 1**	3.81 (0.87)	4.25 (0.64)	3.92 (0.93)	$F(2,62) = 1.57$, $p = .22$
Statement 2**	2.48 (0.87)	2.25 (1.16)	2.46 (1.18)	$F(2,62) = 0.28$, $p = .76$
Statement 3**	3.81 (0.87)	4.10 (0.79)	3.79 (0.78)	$F(2,62) = 0.95$, $p = .39$
Statement 4**	3.71 (0.85)	3.90 (0.79)	3.33 (1.17)	$F(2,62) = 2.01$, $p = .14$
Vocabulary size (LexTALE score)***	71.0 (5.76)	69.9 (7.50)	70.5 (8.36)	$F(2,62) = 0.12$, $p = .89$
Passive knowledge of target words***	7.91 (8.01)	14.79 (15.20)	10.49 (8.87)	$F(2,62) = 2.10$, $p = .13$

Note. Variables marked with one asterisk were self-rated on a 1-5 scale (1 = *very low*, 5 = *very high*). Variables marked with two asterisks were self-rated on a 1-5 scale (1 = *strongly disagree*, 5 = *strongly agree*). Variables marked with three asterisks indicate a percentage. +O versus -O refers to required output production, +NTH versus -NTH refers to noticing the hole.

4.2.2 Materials

The target words were 16 infrequent names of concrete objects that are typically unknown in L2 Dutch for German native speakers. All were non-cognates between Dutch and German, for example *garde* (German: *Schneebesen*, English: *whisk*). There were also 44 filler words that the participants should already have known. These were used to distract from the learning purpose of the study, and because we did not want to present more than one target item in a trial. The fillers were common objects (e.g., an apple). Their cognate status was not controlled. All targets and fillers were depicted through photographs. These had been found on the internet and edited in Photoshop. They were cropped to squared pictures and any words or brand logos were removed.

The complete item list can be found in Appendix A. In the interest of the price judgment cover story, the words came from four semantic categories (children, clothing, household and tools). Each category contained four target words and eleven filler words. Item selection was based on the pre-test data from Chapter 3. Of the target words selected for this study, an average of 1.63% (*SD* = 2.94, range 0–8) was known to the participants (*N* = 32) in Chapter 3; of the fillers 98.40% were known on average (*SD* = 2.18, range = 94–100) (see Appendix A).

In the current study, we did not perform a pre-test on the participants' knowledge of the target words, like we did in Chapter 3. This would have induced NTH in the case of unknown words, which we obviously wanted to avoid in the control condition. Furthermore, the pre-test data from our earlier study showed that the target words were only known to German learners of Dutch in very rare cases, and the filler words were practically always known. Still, all participants in the current study were asked about their pre-existing knowledge of the materials at the end of the experiment (see Debriefing and measures). This allowed us to exclude already-known target words from the analysis.

4.2.3 Procedure

The experiment took place in a quiet laboratory room and lasted 60–75 minutes. The participants received course credit or gift vouchers for their participation. Informed consent was obtained prior to the experiment.

4.2.3.1 Manipulation

NTH was manipulated immediately before the exposure to the target words. The participants were told that the experiment concerned a price judgment task, consisting of two parts: a sorting task and a price comparison task. In the sorting task, the participants were given cards with pictures of the target and filler objects, which should be sorted according to their (subjective) price. The sorting procedure was carried out separately for the objects in each of the four semantic categories. After the participants had finished sorting the first pile of cards, they were given the opportunity to inspect their sorted cards one more time. The participants

had previously been instructed that, in the following price comparison task, they would be required to make price judgments consistent with their self-made ranking.

During this inspection of the cards, the treatment in the experimental and control conditions diverged. The experimental participants were asked to present their ranking to the experimenter by naming, vocally ([+O]) and in Dutch, the objects from the most expensive to the least expensive. We expected them to fail at naming the target objects, thereby experiencing NTH. If a participant did not know what a given object was called, he/she described it in Dutch. Later interviews confirmed that these participants all experienced NTH. In contrast, the control participants were asked to inspect their pile of cards in silence ([–O]), which we expected would not induce NTH. Yet, later interviews showed that this inspection did lead to NTH for about half of the control participants, who were reassigned to a newly created third group for analysis. After they had looked through their cards, the participants commenced the sorting procedure for the next category.

4.2.3.2 Price comparison task

After the sorting task, all participants received naturalistic input containing the target words provided by the experimenter, the author of this thesis and a female native speaker of Dutch. The participant and experimenter were seated opposite each other, each in front of their own keyboard and computer monitor. On these monitors, two objects were displayed per trial, side by side, each picture sized 15x15 cm. As the objects appeared, the experimenter made a statement in Dutch about their relative price, starting with the left object (e.g., “a bed is more expensive than a fridge”). These statements were always reasonable, although not always in accordance with how the participants had previously sorted the cards. The participants were required to press one of two buttons to indicate whether or not the statement agreed with their previously established price ranking. No time limit was imposed for this response, which would not be analysed. Immediately after the response, two new objects appeared on the monitor. The objects always were visible to both the experimenter and participant.

There were four blocks (corresponding to the four semantic categories) with 40 trials per block. The order in which the semantic categories were presented was counterbalanced across the participants, and corresponded to the order of the sorting task. The position of slots for target and filler words in the trial list was fixed, but the assignment of actual target and filler words to slots was random. Each trial contained at most one target. Each target object appeared equally often in the left or right slot. Trials containing targets were always separated by at least one trial with two fillers. Each target object (four per semantic category) was presented four times, with an inter-trial interval of four trials between the first and second, and between the third and fourth exposure. The inter-trial interval between the second and third exposure was 14 trials. The eleven fillers (per category) each appeared five or six times. Each block had a duration of approximately four minutes, and the blocks were separated by a short break.

4.2.3.3 Debriefing and measures

Following the price comparison task, the participants were asked what they thought the study was about. We asked the question at this point to avoid the participants' responses becoming biased by having taken an explicit vocabulary test. After their response, they were told that the experiment was about word learning and they would therefore take a vocabulary test next. This was the first mention of the vocabulary test, which measured immediate learning gains. All objects, including the fillers, were presented successively on the computer screen, in four blocks (in the same order as before). The order of items within the blocks was randomised. The participants were instructed to (try to) name the objects, and received no feedback concerning their response.[2]

After this vocabulary test, the participants filled in a questionnaire about their experience with learning Dutch and other languages. Then, they completed the Dutch version of the LexTALE vocabulary test.

Next, and about 15 minutes after the first vocabulary test, the participants were shown all the target objects (but not the fillers) once more. In this delayed vocabulary test, the participants tried again to name the objects. After each trial, the experimenter provided the correct answer, and asked whether the participant had had passive or active knowledge of the word before taking part in the experiment.

Then, the participants were interviewed to verify whether the NTH manipulation had worked as intended. Initially, we had started by asking the first participants a general question about their experience during the sorting task. However, the participants usually commented on prices rather than on NTH. We then asked them a more specific question (after a while, we stopped asking the first, unspecific question). For the experimental group, the question was: "When naming the pictures after sorting them, did you notice you were not able to produce some names?"

The control group was asked: "When looking at the pictures after sorting them, did you name the objects in silence?". If they said yes, the participants were asked: "In what language?". If they said in Dutch, the participants were asked: "Did you notice you were not able to produce some names?". If participants in the control group said no to the first question, the follow-up questions were not asked. We assumed that not trying to name pictures automatically meant that no NTH took place. We now consider this to be a limitation of the current study, as it would have been better to check this assumption explicitly.

[2] As one reviewer remarked, the immediate post-test would generate NTH in all groups, including [−O, −NTH]. This is inevitable when conducting a vocabulary test, but it confounds the performance on the second post-test of the three groups. However, we do not consider this a problem, because the hypothesised explanation of NTH's facilitative effects rests on how people process the input after noticing holes, and no input was offered in between the two vocabulary tests.

4.2.4 Analysis

4.2.4.1 Reassignment of participants to conditions

As explained earlier, the participants in the control condition were divided into two subgroups for analysis, on the basis of the participants' self-reported experience of NTH. If participants reported that they had subvocally tried to name the objects in Dutch and noticed holes, they were assigned to the [−O, +NTH] group. This includes participants who reported using a combination of Dutch and German for subvocal naming. If participants reported they had exclusively subvocally named the objects in German (their L1) or had not named them at all, they were assigned to the [−O, −NTH] group.

4.2.4.2 Data preparation

The target words of which the participants had reported pre-existing active knowledge were excluded from the analysis.[3] The target words of which the participants had pre-existing passive knowledge were not excluded, because passive knowledge does not preclude word form learning for active production. To take into account any potential effects of passive knowledge on word learning, this information was included in the analysis (see Modelling).

4.2.4.3 Scoring

Learner productions were compared to target productions based on phonological similarity. To this end, we transcribed all learner productions with the DISC phonetic transcription system (Burnage, 1990), which captures every sound of Dutch in one ASCII character, including diphthongs. Details about the phonetic transcription can be found in Appendix B.

Target word responses were scored at the phoneme level. This was preferred to a binary correct/incorrect score, as some word productions were partially correct (e.g., a participant saying *ramlert* to the target *rammelaar*, English: *rattle*). Instead, we counted the number of correctly and incorrectly produced phonemes. Following Levenshtein (1966), deletion, substitution and insertion of phonemes were considered incorrect. In the scoring process we employed long alignment, which lets the same phonemes appear as corresponding segments (see Heeringa, 2004). Table 2 exemplifies the scoring procedure for the *ramlert* example.

[3] To check the reliability of participants' self-reported previous knowledge, we compared the naming data from the participants in the [+O, +NTH] condition, who had named all objects out loud after the sorting task, to their self-reported previous knowledge. These data converged for 99.7%.

Table 2. A target word (*rammelaar*, English: *rattle*) and a participant's production of this word, phonetically transcribed.

Target word	r	ɑ	m	ə	l	a:	r	
Participant's production	r	ɑ	m		l	ə	r	t
Scoring	correct	correct	correct	incorrect (deletion)	correct	incorrect (substitution)	correct	incorrect (insertion)

Ramlert would be counted as yielding five correct and three incorrect phonemes; the corresponding dependent variable for the statistical model for this particular production would in principle be the vector (5,3), representing (Number of correct phonemes, Number of incorrect phonemes). However, the target's actual word length is 7 phonemes. Because we used a binomial probability distribution to predict the number of correct and incorrect phonemes (see Modelling), which does not allow word length to vary within words, we would adjust the final score to be (4,3). A more comprehensive explanation of this issue can be found in Appendix B, but it should be noted that, for 96.4% of the responses, the length of the word produced by the participant was equal to the original word length. For the purpose of providing descriptive statistics, the original vector of correct and incorrect phonemes was also converted into a percentage. This percentage is the number of correct phonemes out of the total number of phonemes (longest alignment). In the *ramlert* example: 5 / (5+3) * 100% = 63%.

4.2.4.4 Modelling

We analysed the data using generalised two-level linear mixed-effects models of the binomial family with the *lme4* package (Bates, Mächler, Bolker & Walker, 2015) in R (R Core Team, 2018). The models were fitted by maximum likelihood estimation, using the logit link function. The vector with the number of correct and incorrect phonemes for each target word utterance was used as the dependent variable. This vector approach to the analysis of proportion data is described in Crawley (2007), and solves four problems that are associated with the alternative of using percentages as a dependent variable (Crawley, 2007, pp. 569–570).

Included as fixed effects were Condition (three levels: [+O, +NTH], [−O, +NTH], [−O, −NTH]), Testing moment (two levels: Immediate, Delayed), and their interactions. As random effects, we included random intercepts for Participant (N = 65) and Word (N = 16). Using this model as a basis, we explored whether its fit to the data could be improved by including random slopes of Testing moment over Participant and Word, which allows for the potential scenario that not all participants or words are equally affected by the 15-minute delay. The results are reported below. We also explored some fixed effects that were not of direct interest to our research questions, but could conceivably affect word learning. These fixed effects were Passive knowledge, the interaction between Passive knowledge and Condition, and Word length (number of phonemes) (Jalbert, Neath, Bireta & Surprenant, 2011). Passive knowledge was a self-reported measurement obtained in the delayed

post-test (see Debriefing and measures). We compared the different nested models using likelihood ratio tests. Alpha was set at .05. Only in case of a significant increase in model fit, in combination with a decrease in the Akaike Information Criterion (AIC; Akaike, 1974), were these additional effects left in the model.

Linear mixed-effects models yield beta estimates relative to the intercept, which represents one specific combination of condition levels. To perform pairwise comparisons across all condition levels, we used the R package *lsmeans* (Lenth, 2016). *lsmeans* uses Tukey's method for *p*-value adjustment in multiple comparisons (Tukey, 1949). As *p*-value adjustment in (generalised) linear mixed-effects models does not seem to be standard practice in the psycholinguistic literature (although it is recommended by Quené & Van den Bergh, 2004), we also provide the unadjusted *p*-values.

4.3 RESULTS

4.3.1 Descriptive statistics

Table 3 shows the mean percentage of correctly produced phonemes over all of the words in the experiment. As was mentioned in the Scoring section (4.2.4.3), these percentages were calculated from the vectors of correct and incorrect phonemes that are used as the dependent variable in our statistical models. In Table 3, the two levels of Passive knowledge (Yes/No) were averaged over.

Table 3. Mean percentage of correctly produced phonemes by Condition and Testing moment, and the correlation between the two testing moments for all conditions.

Condition	Testing moment: Immediate				Testing moment: Delayed (15 min.)				*r*
	Mean	*SD*	95% CI	*n*	Mean	*SD*	95% CI	*n*	
[+O, +NTH]	28.06	11.70	22.73 – 33.38	21	26.03	12.07	20.54 – 31.52	21	0.94
[-O, +NTH]	26.25	12.68	20.31 – 32.18	20	23.13	13.66	16.73 – 29.52	20	0.92
[-O, -NTH]	16.54	10.91	11.93 – 21.15	24	16.52	11.22	11.78 – 21.26	24	0.89
Total	23.25	12.67	20.11 – 26.39	65	21.63	12.77	18.46 – 24.79	65	0.92

Note. n indicates the number of participants in each condition.

To ease interpretation, Table 4 shows what percentage of target words were actually produced correctly, partially correctly, and incorrectly.

Table 4. Percentage of words that were produced fully correctly, partially correctly, and fully incorrectly (by Condition and Testing moment).

Condition	Testing moment: Immediate			Testing moment: Delayed (15 min.)		
	Correct	Partial	Incorrect	Correct	Partial	Incorrect
[+O, +NTH]	19%	15%	66%	18%	13%	69%
[-O, +NTH]	15%	18%	67%	14%	15%	71%
[-O, -NTH]	11%	9%	80%	10%	10%	80%
Total	15%	14%	71%	14%	12%	74%

Table 5 is similar to Table 3, but here the scores are divided by Passive knowledge (Yes/No) rather than by Testing moment (which is now averaged over).

Table 5. Mean percentage of correctly produced phonemes by Condition and Passive knowledge.

Condition	Passive knowledge: No				Passive knowledge: Yes			
	Mean	*SD*	95% CI	*n*	Mean	*SD*	95% CI	*n*
[+O, +NTH]	25.47	10.39	20.74 – 30.20	92.09	46.88	33.86	25.36 – 68.39	7.91
[-O, +NTH]	22.89	14.43	16.14 – 29.64	85.21	28.84	29.30	12.61 – 45.07	14.79
[-O, -NTH]	16.65	11.98	11.59 – 21.71	89.51	14.74	27.62	1.00 – 28.47	10.49
Total	21.42	12.72	18.27 – 24.57	89.35	28.01	32.00	18.39 – 37.62	10.65

Note. n indicates the mean percentage of items that were passively known or unknown in each condition.

In the following, we will report the inferential statistics that tell us whether or not the contrasts shown in these tables reached significance. Before doing so, we will report the model comparisons leading up to the final model we used to arrive at the inferential statistics.

4.3.2 Model comparisons

The inclusion of a random slope of Testing moment over Participant did not significantly improve model fit ($\chi^2 = 4.12$, $df = 2$, $p = .13$). Another non-significant result was found for the random slope of Testing moment over Word ($\chi^2 = 0.62$, $df = 2$, $p = .73$). Thus, these random effects were not included in the final model.

We then explored the fixed effects. Passive knowledge significantly increased model fit ($\chi^2 = 21.64$, $df = 1$, $p < .001$, AIC decreased from 7522.9 to 7503.2), as did the subsequent addition of its interaction with Condition ($\chi^2 = 34.58$, $df = 2$, $p < .001$, AIC decreased from 7503.2 to 7472.6). Word length did not improve model fit ($\chi^2 = 1.91$, $df = 1$, $p = .17$), and was again removed from the model.

Thus, the final model was specified as follows: (Number of correct phonemes, Number of incorrect phonemes) ~ 1 + Condition * Testing moment + Condition * Passive knowledge + (1 | Participant) + (1 | Word). In this notation, the dependent variable on the left of the '~' is

modelled from the fixed and random effects on the right of the '~', '1' represents an intercept, '*' represents an interaction including all lower-order effects, and '|' indicates random effects.

4.3.3 Inferential statistics

The estimates of our generalised linear mixed-effects model are shown in Table 6. These estimates are approximations of the binomial parameter, which here concerns the probability that a phoneme is produced correctly. The estimates are given on the logit scale, and can be back-transformed to probabilities with the formula $e^x / (1+e^x)$, where x is the logit. To obtain the logit for a specific combination of variable levels that is not the intercept, for example for [+O, +NTH] at delayed testing with no pre-existing passive knowledge, one should add the corresponding logit estimates to that of the intercept (in this example: −2.29 + 0.85 – 0.04 – 0.10 = −1.58).

Table 6. Model outcomes.

Fixed effects	Logit	Odds ratio	*SE*	*z*	*p*
(Intercept)	-2.29	0.10	0.41	-5.64	**< .001**
Condition: [+O, +NTH]	0.85	2.34	0.31	2.78	**.005**
Condition: [-O, +NTH]	0.62	1.86	0.31	1.98	**.048**
Testing moment: Delayed	-0.04	0.96	0.10	-0.46	.64
Condition: [+O, +NTH] and Testing moment: Delayed	-0.10	0.90	0.13	-0.73	.47
Condition: [-O, +NTH] and Testing moment: Delayed	-0.17	0.84	0.14	-1.23	.22
Passive knowledge: Yes	-0.34	0.71	0.17	-1.99	**.047**
Condition: [+O, +NTH] and Passive knowledge: Yes	1.32	3.74	0.24	5.48	**< .001**
Condition: [-O, +NTH] and Passive knowledge: Yes	1.00	2.72	0.23	4.41	**< .001**
Random effects	**Variance**	***SD***			
Participant (intercept)	0.94	0.97			
Word (intercept)	1.90	1.38			

Note. The intercept represents the following combination of variable levels: Condition = [−O, −NTH], Testing moment = Immediate, and Passive knowledge = No. Significant *p*-values are printed in bold.

The odds ratio is a measurement of effect size. With the exception of the intercept itself, these numbers show how the odds of correctly producing a phoneme change for a specific level of a variable, as compared to the level represented by intercept. For example, for participants in the [+O, +NTH] group, the odds to correctly produce a phoneme are estimated to be 2.34 times higher than for participants in the [−O, −NTH] group[4] (at immediate testing and with no pre-existing passive knowledge, see the paragraph below).

In mixed-effects models, the intercept always represents one specific combination of variable levels. Here, it represents the [−O, −NTH] group, tested immediately, and on words for which no pre-existing knowledge was reported. From Table 6, it can be seen that [−O,

−NTH] under these circumstances was significantly outperformed by [+O, +NTH] (p = .005) and by [−O, +NTH] (p = .048). However, Table 6 alone does not inform us on contrasts that do not involve the intercept (for example, if we wanted to contrast [−O, +NTH] with [+O, +NTH]). Using the *lsmeans* package (Lenth, 2016), the data have been rearranged in Table 7 to show pairwise comparisons for all Condition contrasts at both testing moments. For simplicity, the levels of Passive knowledge are averaged over. This explains why the first two odds ratios in Table 7 (4.57 and 3.06) are not the same as those reported in Table 6 (2.34 and 1.86), which only applied to Passive knowledge = No. As can be seen, the correction of p-values for multiple testing does not change the significance of the findings.

Table 7. Pairwise comparisons among the estimated means for all con-ditions, averaged over Passive knowledge (Yes/No).

Testing moment	Contrast	Logit	Odds ratio	*SE*	*z*	Unadjusted *p*	Adjusted *p*
Immediate	[+O, +NTH] – [-O, -NTH]	1.52	4.57	0.32	4.71	**< .001**	**< .001**
	[-O, +NTH] – [-O, -NTH]	1.12	3.06	0.32	3.48	**< .001**	**.002**
	[+O, +NTH] – [-O, +NTH]	0.40	1.49	0.33	1.21	.23	.45
Delayed	[+O, +NTH] – [-O, -NTH]	1.42	4.14	0.32	4.41	**< .001**	**< .001**
	[-O, +NTH] – [-O, -NTH]	0.95	2.59	0.32	2.96	**.003**	**.009**
	[+O, +NTH] – [-O, +NTH]	0.47	1.60	0.33	1.42	.15	.33

Note. Significant p-values are printed in bold.

The pairwise comparisons tell us that participants in the [+O, +NTH] group scored significantly higher than participants in the [−O, −NTH] group, both at immediate testing (p < .001) and after a 15-minute delay (p < .001). Both odds ratios (immediate: 4.57, delayed: 4.14) can be considered of approximately medium magnitude. More concretely, as can be calculated from Table 3, at immediate testing, the number of correctly produced phonemes was 70% higher in the [+O, +NTH] group than the [−O, −NTH] group. After 15 minutes, the [+O, +NTH] participants still produced 58% more correct phonemes as compared to the [−O, −NTH] participants.

The [−O, +NTH] participants also outperformed their peers in the [−O, −NTH] group, both at immediate testing (p = .002) and at delayed testing (p = .009). These effect sizes (immediate: 3.06, delayed: 2.59) were smaller. Still, the participants in the [−O, +NTH] group produced 59% more phonemes correctly at immediate testing, and 40% after 15 minutes, as compared to their peers in the [−O, −NTH] group. Finally, no significant difference could be detected

[4] Unfortunately, for L2 research, no standardised guidelines for the interpretation of odds ratios exist. Different guidelines that are available suggest that 1.5/2.5/4.3 (The Effect Size, n.d.), 1.5/3.5/9 (Hopkins, 2002), or 1.68/3.47/6.71 (Chen, Henian & Chen, 2010) can be considered as small/medium/large.

between participants in the [+O, +NTH] and the [−O, +NTH] groups, who had both noticed holes (p = .45 at immediate testing, and p = .33 at delayed testing).

With regard to Testing moment (see Table 6 again), there was no significant decay over a period of 15 minutes time (p = .64). Table 8 shows that the interaction between Testing moment and Condition was not significant for any of the contrasts (all adjusted ps > .44).

Table 8. Pairwise comparisons of the interaction between Condition and Testing moment.

Contrast	Logit	Odds ratio	SE	*z*	Unadjusted *p*	Adjusted *p*
[+O, +NTH] – [-O, -NTH]	-0.10	0.90	0.13	-0.73	.47	.75
[-O, +NTH] – [-O, -NTH]	-0.17	0.84	0.14	-1.23	.22	.44
[+O, +NTH] – [-O, +NTH]	0.07	1.07	0.13	0.54	.59	.85

Finally, Table 6 shows an interaction between Condition and Passive knowledge. Pre-existing passive knowledge had a negative effect on the learning rate for participants in the [−O, −NTH] group (p = .047). The odds ratio was 0.71, which means that these participants were 1.41 (= 1/0.71) times more likely to correctly produce a phoneme in a word they had had no pre-existing knowledge of than a phoneme in a word they had had pre-existing knowledge of. In the participants who noticed holes, pre-existing passive knowledge had a larger and positive effect (in [+O, +NTH]: p < .001, OR = 2.67, and in [−O, +NTH]: p < .001, OR = 1.93; these estimates were obtained through relevelling).

4.4 DISCUSSION

In this study, we asked whether NTH (i.e., the awareness of vocabulary holes or gaps) in spoken L2 word production facilitates the acquisition of these words from subsequent spoken input in an incidental learning environment. We created this environment by conducting the experiment outside of the classroom, and in the country where the target language was spoken. The incidental aspect of the study is also reflected by the fact that none of the 65 participants in the final sample suspected that the experiment was a language learning study, as we verified in post-experiment interviews.

4.4.1 From two to three conditions

The original design included two conditions. In the experimental condition, the participants were required to vocally produce the target words. Because they did not actually know these target words, they failed in producing them, and thereby noticed holes in their vocabulary. Thus, *output* in the current study does not refer to language production in the typical sense, but rather to the requirement of output. The experimental participants then were exposed to input containing the unknown vocabulary.

In the control condition, the participants studied pictures without being asked to name them and therefore were supposed not to notice holes. Then, they were exposed to

the same input as the experimental group. However, about half of the participants in the control condition indicated they had subvocally tried to name (some of the) objects in L2 Dutch. Although we did not explicitly ask them whether these subvocal naming attempts had resulted in the experience of NTH (which we consider a limitation of the current study), it does seem very likely that this was the case. In other words, these participants should have experienced what Godfroid, Housen and Boers (2010) call "learner-induced noticing" (also see Park, 2007; Williams, 1999). Given this situation, we divided the control condition into two new groups for analysis: [−O, +NTH] and [−O, −NTH]. The experimental condition was renamed [+O, +NTH].

Following Festinger's (1957) theory of cognitive dissonance, one might wonder whether the self-reported (absence of) NTH in the control participants was influenced by their post-test performance. In other words, did the participants who learned fewer words perhaps 'justify' this outcome by claiming that they had not named the objects in Dutch in the sorting task? This seems unlikely: Sorting the cards took place before the participants were exposed to input, and thus bore no obvious relationship to the effort that the participants made to learn words. Indeed, during the interviews, the participants did not show any evidence of associating particular sorting strategies with particular word learning outcomes.

We also compared the three groups on eleven variables related to word learning (see Table 1), and no significant differences were found. In the context of this study, this is a good thing: The conclusions we have drawn from our analysis should not have been biased by group-level differences in one or more of these variables. At the same time, it means that we still do not know what caused some control participants, but not others, to notice holes. The individual differences that would explain why some people are more likely than others to experience learner-induced noticing are something to be explored further in future research.

4.4.2 Effect of NTH, and underlying mechanisms

We will now consider our main research question concerning the effect of NTH on L2 word learning from spoken input. The results showed that NTH facilitates word learning, which is in line with Swain's hypothesis on the noticing function of output (1985, 1993, 1995, 1998). The effect was found both when NTH was experimentally induced by requiring the participants to produce output, and when it was not induced through required output but still internally generated by the participants. Swain (1995, p. 125) already mentioned in passing that (failure in) language production may be vocal or subvocal for the noticing function of output to have an effect. We believe to be the first to have empirically demonstrated this, through the finding that the [+O, +NTH] and [−O, +NTH] participants both outperformed the [−O, −NTH] participants. For the strength of the effect it did not matter whether vocal language production was required or was not required but happened subvocally: [+O, +NTH] and [−O, +NTH] were not significantly different from one another.

The benefit of noticing holes on L2 word learning can potentially be explained by the mechanisms that were mentioned in the Introduction. These mechanisms can be

summarised as learners allocating more attentional resources to the input after having become aware of their linguistic problems or vocabulary holes, and being curious as to how to resolve or fill those. Perhaps NTH functions as a type of orienting, which is one of three major attentional systems proposed by Posner & Petersen (1990). This system commits attentional resources to sensory stimuli (Tomlin & Villa, 1994, p. 190). Since our explanation for the effect of NTH rests on how the participants processed the input after having noticed holes, it is understandable that it did not matter whether NTH took place with or without (an attempt to) vocal output production.

4.4.2.1 Suggested direction for future research: Mediation analysis

The mechanisms discussed in the above paragraph could be empirically investigated in future studies using mediation analysis (see Imai, Keele, Tingley & Yamamoto, 2011; MacKinnon, Fairchild & Fritz, 2007). Finding empirical support for such hypothesised pathways would mean a great step forward in our understanding of exactly how the positive relationship between NTH and L2 word learning comes about. It is almost certain that at least one further variable must be involved (and potentially more). After all, the realisation of a vocabulary hole in itself does not fill up that hole with the right word form. Rather, an explanation based on mediation through a third and fourth variable was already given: Experiencing NTH could make learners curious about the word forms missing in their vocabulary, which in turn could lead them to allocate more attention to the input, leading to more word learning.

Thus, we propose the following chain of processes: NTH → curiosity → attention → word learning (while recognising that this chain is not necessarily exhaustive, and that alternative chains could exist as well). In order to investigate this chain, a future study should also measure curiosity and the amount of attention paid to the target vocabulary during the price comparison task. Attention might be measured using eye tracking (e.g., Godfroid, Boers & Housen, 2013). Curiosity could potentially be measured in a stimulated recall procedure (Gass & Mackey, 2000) after the task is finished. If participants were questioned regarding their curiosity about learning words before or during exposure to input, this would likely trigger NTH in participants assigned to conditions in which no NTH should take place.

In the current study, the incidental finding that some participants in the original control condition had noticed holes enabled us to make some additional comparisons between the groups that we had not initially foreseen. For studying the noticing function of output, this was very interesting. If one wanted to conduct a mediation analysis, however, it would be necessary to have access to a manipulation of NTH that works predictably for all participants. Specifically, participants in a control group should not experience NTH. One potential solution for the current set-up could be to leave out the sorting task for the control group. Then, the control participants would not experience NTH before being exposed to input. This would have the disadvantage, however, that participants in the [+NTH] group would already be more familiar with the materials at the start of the price comparison task.

Alternatively, mediation analysis can also be applied to studies in which participants are assigned to conditions based on their self-reported experience of NTH, as we did in the current study. However, a prerequisite is that we would need to know the variable(s) that lead some learners but not others in the [–O] condition to experience NTH (Imai et al., 2011). The variables we included in Table 1 did not explain this difference, so further exploration would be required. A disadvantage of applying mediation analysis to a study using non-random assignment is that one cannot be sure whether a significant third variable actually is a mediator variable, rather than a confounding variable. In the latter case, the third variable would both cause learners to experience NTH on the one hand, and on the other hand to be more curious or to allocate more attention to language. The question of mediation versus confounding can be resolved if a predictable method for manipulating NTH is found: Only if the third variable is a mediator and not a confound, a relationship between the independent variable (NTH) and the mediator should become visible upon manipulating the independent variable.

4.4.3 Effect of Testing moment

Another question of this study was at what rate newly-acquired L2 word knowledge is again forgotten. We found no significant decrease in scores over a period of 15 minutes (although a trend towards decay was visible). Thus, it seems that Ebbinghaus's (1885/1913/2011) nonsense syllables were forgotten sooner (he only remembered 58% after 20 minutes) than the L2 vocabulary in this experiment. Of course, learning a list of nonsense syllables is not the same as learning meaningful L2 names of real objects. Potentially, the current participants had a higher motivation to remember the vocabulary they had just learned, or benefited from the connection that could be made between the word forms and their object referents.

Perhaps due to the short delay of 15 minutes, there was no significant interaction effect between Condition and Testing moment either: Word knowledge did not decay at different rates depending on the condition. Thus, the differences between the conditions that were observed at immediate testing persisted 15 minutes later. Readers interested in the retention of word knowledge over longer periods of time are referred to Chapter 3. That study did show a significant decline in word knowledge in tests after both 20 minutes and six months following exposure. However, that study was different from the current study in several aspects. In conclusion, the retention of incidentally acquired L2 word knowledge over short periods of time seems to depend on the task in which this knowledge was acquired.

4.4.4 Effect of Passive knowledge

Because we worked with natural language items, there was the possibility that the participants would already have (some) knowledge of the target words before commencing the experiment (even though we had pre-tested all our items on a similar participant group, see Materials). This was checked through self-report at the end of the experiment. Words

that a participant already had actively known before taking part were removed from the analysis. Words of which only passive knowledge was reported were included in the analysis, and we investigated whether such pre-existing passive knowledge was related to learning success on the word level.

The participants who had noticed holes (with or without required output) achieved significantly higher learning scores on those words they had already had passive knowledge of. For the participants who had not noticed holes, the relationship was the other way around: They achieved significantly lower learning scores on words they had already passively known before. While this initially may seem surprising, an explanation is conceivable.

The participants had not been told that they would be tested on object names in a picture-naming post-test. Thus, when they were exposed to the target words in the price comparison task, they probably were not consciously preparing themselves for such a task (please recall that, in contrast to Chapter 3, the participants in the current chapter did not produce output during the price judgment task). Since it is known that people generally pay more attention to novel stimuli (e.g., Horstmann & Herwig, 2016; Johnston, Hawley, Plewe, Elliott & Dewitt, 1990), it is likely that the participants paid more attention to the target words they had never heard before, and, as a result, better acquired those word forms. This could explain the (weak) negative effect of pre-existing passive knowledge on word learning for the participants who had not noticed holes.

Why would this not apply to the participants who had noticed holes in the sorting task? Their passive knowledge was of no use in the moment when they had to retrieve the names of the target objects from memory. Thus, these participants experienced NTH for all the target objects they could not name, regardless of whether or not they knew their names passively. This also means that they presumably became curious about all of these names, again regardless of passive knowledge status. Then, in the price comparison task, the participants probably paid extra attention to all the objects they were unable to name before. Upon hearing these objects' names, the participants' already existing knowledge of these names was reactivated, and there was less new information to be learned. This could explain the positive relationship between pre-existing passive knowledge and word learning in the participants who had noticed holes, and why the directionality of the relationship differed between participants who had and had not noticed holes.

4.5 SUMMARY AND CONCLUSIONS

This study showed that noticing holes in one's vocabulary facilitates subsequent incidental L2 word learning from spoken input. Participants who reported awareness of not being able to produce certain words acquired more of these words from later input, as compared to participants who did not report such awareness. It did not matter whether this awareness had been experimentally induced by requiring the participants to vocally produce output (and fail), or whether it was learner-generated and resulted from subvocal (failure in) output production. The current study does not yet allow us to also draw conclusions about

the cognitive mechanisms that explain the increase in word learning rates following the experience of NTH. Therefore, we suggest that future researchers use mediation analysis to explore the mechanisms that underlie the effects of Swain's noticing function of output.

In addition to these theoretical insights, there are two practical lessons to be drawn from this study. Firstly, when it comes to studying NTH (and presumably other forms of noticing too), even under identical treatment conditions participants can differ in their actual NTH experience. This means that NTH will always need to be monitored, rather than just assumed to be present or absent. Secondly, although for word learning it did not matter whether NTH was induced by pushing the learners to produce output or was learner-generated, only the pushed-output treatment generated NTH for all participants in the first place. Thus, if language teachers wanted their students to experience NTH, pushing them to produce output seems worthwhile.

In conclusion, when learners become aware of their vocabulary holes, the first step in filling these holes is already taken. The fact that these results were found in a setting that did not explicitly encourage participants to learn words is very promising. Conceivably, in classroom contexts focused on language learning, effects of NTH might be even more pronounced. This should be investigated in future studies: Such knowledge would be very relevant to both language teachers and learners.

APPENDIX A: ITEMS

Table A contains all the target items that were used in this experiment, and indicates to what extent they were already actively known in a comparable participant population. Table B contains the same information for the filler items.

Table A. Target items.

Dutch	German	English	Category	% known
rammelaar	Rassel	rattle	children	0
romper	Body	onesie	children	6
sambabal	Maraca/Rumba-Rassel	maraca	children	0
tol	Kreisel	top	children	0
gesp	Gürtelschnalle	clasp	clothes	0
kous	Strumpf	stocking	clothes	8
slab	Lätzchen	bib	clothes	0
tooi	Federschmuck	headdress	clothes	0
garde	Schneebesen	whisk	household	0
lessenaar	Pult	lectern	household	0
stolp	Glasglocke	(bell-)glass	household	0
waaier	Fächer	fan	household	0
dobber	Schwimmer	float	tools	0
klos	Rolle	reel (of cotton)	tools	6
passer	Zirkel	compass	tools	0
vijzel	Mörser	mortar	tools	6

Note. "% known" indicates how many participants (N = 32) in Chapter 3 could name this word in a picture naming pre-test.

Table B. Filler items.

Dutch	German	English	Category	% known
appel	Apfel	apple	children	100
bal	Ball	ball	children	97
banaan	Banane	banana	children	97
boek	Buch	book	children	100
gameboy	Gameboy	game boy	children	97
hond	Hund	dog	children	100
kat	Katze	cat	children	100
muffin	Muffin	muffin	children	100
paard	Pferd	horse	children	100
skateboard	Skateboard	skateboard	children	97

vogel	Vogel	bird	children	97
beha	BH	bra	clothes	94
bikini	Bikini	bikini	clothes	97
bril	Brille	glasses	clothes	90
handdoek	Handtuch	towel	clothes	97
parfum	Parfum	perfume	clothes	94
ring	Ring	ring	clothes	100
schoen	Schuh	shoe	clothes	97
sjaal	Schal	scarf	clothes	97
sok	Socke	sock	clothes	97
tas	Tasche	bag	clothes	97
t-shirt	T-Shirt	t-shirt	clothes	100
bed	Bett	bed	household	100
deur	Tür	door	household	100
koelkast	Kühlschrank	fridge	household	97
lamp	Lampe	lamp	household	100
pan	Pfanne	pan	household	97
plant	Pflanze	plant	household	97
radio	Radio	radio	household	100
sleutel	Schlüssel	key	household	100
spiegel	Spiegel	mirror	household	100
stoel	Stuhl	chair	household	100
wasmachine	Waschmaschine	washing machine	household	97
auto	Auto	car	tools	100
bus	Bus	bus	tools	100
cd	CD	CD	tools	100
fiets	Fahrrad	bicycle	tools	100
laptop	Laptop	laptop	tools	100
microfoon	Mikrophon	microphone	tools	100
smartphone	Handy	smartphone	tools	100
telefoon	Telefon	telephone	tools	100
toilet	Toilette	toilet	tools	100
trein	Zug	train	tools	97
tv	TV	television	tools	100

Note. "% known" indicates how many participants (N = 32) in Chapter 3 could name this word in a picture naming pre-test.

APPENDIX B: SCORING DETAILS

The following exceptions were made in the literal transcription of participants' word productions:

1. Because of their omnipresence and productivity in Dutch (Shetter, 1959), adding a diminutive suffix to a noun was not regarded as insertion (e.g. *slabje* for *slab*, English: *bib*).
2. If participants modified their production in such a way that it was clear that they were only expressing insecurity with regard to their utterance (example: *samba-iets*, literally in English: *maraca something*) and not an actual memory representation, this modification (*iets*, English: *something*) was not regarded as insertion.
3. Sometimes participants gave multiple productions for one word (e.g. "Is it a *stomp*? Or a *stolp*?" for *stolp*, English: *bell jar*). In these cases, the last production was transcribed.
4. If a participant mispronounced a phoneme just because of their German-accented Dutch, it was not marked incorrect (an accent does not reflect a false memory representation).
5. In Dutch, syllable-final obstruents get devoiced. Therefore, upon hearing the word /slɑp/ (*slab*, English: *bib*), one cannot know whether the true underlying form of the final consonant is /p/ or /b/. Therefore, productions such as /slɑpər/ or /slɑbər/ would receive the same score.

Now, we will further discuss the word length issue mentioned under Scoring. In the *ramlert* example, a consequence of using the long alignment is that the sum of correct and incorrect phonemes amounts to 8. However, for participants who did not produce any insertions, the total number of phonemes would be 7 (i.e., equal to the word length). In 3.6% of the data points, the sum of the number of scored phonemes was larger than the total word length. This is problematic, as the binomial probability distribution for a particular word is characterised by a fixed parameter *N* for the number of trials (i.e., the number of phonemes), which should not vary over participants. We resolved this issue by rescaling the number of correct and incorrect phonemes, so that they would always add up to the (fixed) word length of the target word, in this case 7. Rescaling was done by multiplying the word length of the target word by the percentage correct (e.g. 7 * 0.63 = 4.38, rounded off as 4), and subtracting this number from the total number of phonemes to arrive at the rescaled number of incorrect phonemes (7 – 4 = 3). Thus, the final vector for *ramlert* would be (4,3).

5.

Studying in Dutch or English: Does it affect language development?

This chapter is based on:

De Vos, J. F., Schriefers, H., & Lemhöfer, K. (2019).
Studying in Dutch or English: Does it affect language development?
Manuscript in preparation.

ABSTRACT

Nowadays, many study programmes in the Netherlands are offered in English. In the Dutch media, it is speculated that there is a relationship between the language in which students are instructed (English or Dutch), and their language skills in that particular language. For example, when students are instructed in English, this would be beneficial for their English language skills. But there is little empirical evidence for such claims. Therefore, we tracked the language development of 315 Dutch and German students in Nijmegen, the Netherlands. They all studied psychology, either in a Dutch or in an English track. We examined the students' lexical richness (i.e., their productive vocabulary knowledge) at three moments in time during the first year of study. Averaged over these three moments, there seemed to be a native language advantage in lexical richness: The Dutch lexical richness scores of Dutch students who studied in Dutch were generally higher than the English lexical richness scores of Dutch and German students who studied in English. However, they were not higher than the Dutch lexical richness scores of German students who studied in Dutch. We did not detect any evidence that students' lexical richness in Dutch and English would develop at a different speed. This held both for Dutch and German students. Thus, our data suggest that if students want to improve their language skills, the benefits of choosing to study in one language are not greater than choosing to study in the other.

5.1 INTRODUCTION

In the academic year 2016-2017, 69% of all master's programmes at Dutch universities were fully taught in English, as well as 20% of all bachelor's programmes (KNAW, 2017). Over one third of all students was enrolled in such a programme (KNAW, 2017), and the use of English in Dutch higher education is only increasing. In fact, figures of the academic year 2012-2013 showed that no other continental European country offered as many higher education programmes taught in English as the Netherlands (Wächter & Maiworm, 2014). Given these developments, Dutch policy makers are interested in knowing more about the potential effects of English-medium instruction (EMI) on language development and study success of students in the Netherlands. In using the term *EMI*, we follow Macaro, Curle, Pun, An and Dearden (2018, p. 37), who define it as "the use of the English language to teach academic subjects (other than English itself) in countries or jurisdictions where the first language of the majority of the population is not English."

In 2016, the Dutch minister of Education requested that the Royal Netherlands Academy of Arts and Sciences (Dutch abbreviation: KNAW) investigate and review the effects of the use of English in higher education. The resulting report was published in 2017 and, among other things, summarised the different arguments that can explain the advance of EMI at Dutch universities. There are "business-related" arguments (KNAW, 2017, p. 11), such as attracting international students and staff. Furthermore, it is often believed that so-called *international classrooms* increase the quality of education. Then, there are arguments that relate to the labour market. The report states that study programmes that prepare students for the Dutch job market would likely opt for Dutch as the medium of instruction, whereas English would be the preferred instruction language when preparing students for the international job market (KNAW, 2017).

That last argument for choosing English or Dutch as the language of instruction, namely to improve students' proficiency in the respective language, sounds relatively uncontroversial. However, the available literature does not immediately confirm these assumed positive effects of study language on language proficiency. Macaro et al. (2018) conducted a systematic review to evaluate this and other arguments related to EMI. They found 83 empirical studies which investigated EMI in higher education in countries where the majority of the population's first language (L1) was not English. Only seven out of those studies looked at the effect of EMI on the development of English as a second language (L2). This, combined with the diversity in test types, makes it "extremely difficult" to properly assess this issue (Macaro et al., 2018, p. 57).

According to Macaro et al. (2018), some studies found that the L2 English proficiency of students significantly increased over some set period of time on some (but not necessarily all) proficiency measures (e.g., Aguilar & Muñoz, 2014; Rogier, 2012; Yang, 2015). However, Macaro et al. (2018) point out that none of these three studies included a control group. Thus, we do not know whether the increase in proficiency was really and exclusively due to EMI. Lei and Hu (2014) did employ a control group, and found no significant effect of EMI after

partialling out pre-existing differences between the groups. In the end, Macaro et al. (2018) state that the findings are inconclusive and that more research is needed.

There is more research that was not part of the review by Macaro et al. (2018). In a cross-sectional study, Baumgarten (2014) found that L1 speakers of Danish or German were no better at producing English recurrent multiword sequences (for example: "I don't know if", p. 8) in the third as compared to the first year of a trilingual (English, German, Danish) undergraduate programme. Again, this study did not include a control group, but given the fact that even the 'treatment group' with EMI did not improve, the conclusion of no improvement would probably have been the same with or without a control group.

A longitudinal line of research revolved around students at an Australian university who spoke English as an L2. Thus, in contrast to the studies discussed above, the participants were also living in the country where the L2 was spoken. Storch (2009) found that over the course of one semester, the students' English essays improved in structure and ideas, but not linguistic accuracy. One year (Knoch, Rouhshad & Storch, 2014) and three years (Knoch, Rouhshad, Oon & Storch, 2015) after study onset, the students' writings had only improved in fluency, but not in accuracy, grammatical and lexical complexity, and global writing scores. Only a minority of the participants in the three studies were also enrolled in language classes. The findings from these studies show that simply studying for a degree in an L2, even in combination with living in the country where that L2 is spoken, does not necessarily suffice for increasing various aspects of L2 (writing) proficiency.

In conclusion, it seems that the benefits of EMI on L2 English language development are either small or non-existent, but more evidence is needed. To our knowledge, no studies have examined this issue in the Dutch context, or have focused on the development of Dutch proficiency of university students (although some studies have looked at students' Dutch language skills at a fixed point in time, for example Van Houtven, Peters & El Morabit, 2010). The 2017 KNAW report (pp. 63-65) only contains qualitative findings based on interviews with lecturers and students, but no empirical data on language development. In contrast, in this chapter we present empirical data regarding the development of Dutch and English language proficiency during the first year of study at university.

5.1.1 The present study

The aims of this study were to investigate the effect of study language on language development (this chapter) and on study success (e.g., grades; next chapter) of students at a Dutch university. We had access to data of 675 first-year psychology students at Radboud University in Nijmegen, the Netherlands. From the academic year 2016-2017 onwards, the Psychology programme has been offered in two tracks: a fully English one (which I will call *the English track*), and a track in which classes are taught in Dutch, while most study materials are still in English (*the Dutch track*). This situation is perfect for studying the effects of teaching in English versus Dutch, because the curriculum is exactly the same except for the language in which the classes are taught. The lectures are even given by the same, native

Dutch lecturers, who deliver them either in Dutch or in English (although the accompanying work groups may be taught by different teachers).

Psychology at Radboud University attracts a high number of German students, especially since the English track has become available. We therefore were in a position to also compare the outcomes of Dutch and German students. This is interesting because both Dutch and English are an L2 to German students, but German students in the Dutch track are presumably immersed in their study language (because they study in the Netherlands), while German and Dutch students in the English track are not. This enabled us to look at potential immersion effects. Thus, we made comparisons between four groups: Dutch students in the Dutch track, Dutch students in the English track, German students in the Dutch track, and German students in the English track. Both Dutch and German students have had many years of English education by the time they go to university, and should be able to hold conversations and read books in English.

The Psychology programme provided us with data to work with. These consisted of basic demographic information (including age and nationality), various measures of study success (e.g., grades), and three writing samples per student (from open questions of written exams). The data are described in more detail in the Methods section (5.2.3). It was not possible to collect any measures of our own. This imposed some limits on the conclusions we could draw, as will be discussed throughout this chapter.

5.1.2 Investigating the development of lexical richness

We investigated language development by looking at students' written answers to open exam questions at three points in time in their first year of study. Since this thesis concerns L2 word learning, we specifically focused on the students' lexical development. Lexical development was operationalised by means of lexical richness, because this can be computed directly from writing samples and does not require additional vocabulary tests. Lexical richness reflects the sophistication and range of someone's productive vocabulary (Wolfe-Quintero, Inagaki & Kim, 1998, cited in Lu, 2012). Read (2000) has conceptualised lexical richness as existing of four dimensions.

The first dimension, lexical density, concerns the ratio between the number of content words (i.e., nouns, verbs, adjectives and adverbs) and the total number of words in a text. Texts that contain many function words and few content words are therefore considered less lexically 'dense'. The second dimension, lexical sophistication, concerns the ratio between the number of 'sophisticated' words (i.e., relatively difficult or rare words) and the total number of words in a text. Thus, lexically sophisticated texts contain a high percentage of sophisticated words. The third dimension, lexical variation, concerns the diversity of the vocabulary that is used, for example the number of different words relative to the total number of words. Thus, texts with many repetitions of the same words would get lower scores on lexical variation. The fourth and last dimension concerns the number of lexical errors in vocabulary use (i.e., wrong choices of words). In this study, we considered only the first three dimensions because

they could be automatically computed from the written samples with existing software (Lu, 2013).

Lu (2012) has collected a list of 26 different measures that have been defined by researchers to capture the first three dimensions of lexical richness. He investigated the correlation of these 26 measures with ratings of oral L2 English proficiency of Mandarin native speakers, made by English teachers. Lu found significant correlations for some, but not all of the measures. For this study, we selected three measures to work with out of the 26 available measures, one measure per dimension. We based our choice of measures on conceptual considerations that are discussed in the Methods section (5.2.3.3). As a methodological check, we wished to verify whether our three selected measures could distinguish between more and less proficient language users. To this end, the first question we assessed was whether the native speakers in our sample obtained the highest scores. Thus, we expected to see higher scores for Dutch students in the Dutch track on lexical density, sophistication and variation as compared to the three other groups.

Next, we investigated whether EMI is beneficial for the development of English language skills. Since the students all answered the same exam questions, we examined whether the development of (Dutch or English) lexical richness over time was different for students in the Dutch and English tracks. To begin with, we compared the development of lexical richness scores of Dutch students in the Dutch and English tracks. We expected to find more development in L2 English (i.e., in the English track) than in L1 Dutch (i.e., in the Dutch track), because native Dutch speakers have been exposed to Dutch for many years already and thus there might be fewer new Dutch words for them to learn. If this is indeed true, we should see an interaction between the time of measurement (i.e., the first versus second versus third exam), and group (i.e., Dutch students in the Dutch track versus Dutch students in the English track). Such an interaction would support the claim of beneficial effects of EMI: The impact of study language on language skills would be relatively greater in English than in Dutch.

We made a similar comparison between the German students in the two tracks. This enabled us to compare L2 lexical development between learners who, presumably, were immersed in the L2 environment outside their study context (i.e., the German students in the Dutch track), and those who were not (i.e., the German students in the English track). We expected the immersion of German students in Nijmegen's Dutch language environment to have a positive effect on their Dutch lexical development. Therefore, we hypothesised to see more development of Dutch language skills for German students in the Dutch track as compared to development of English language skills for German students in the English track. Again, this should be reflected as an interaction between the time of measurement, and group.

5.1.3 Research questions

In summary, Chapter 5 addressed the following three questions:

1. Is lexical richness the highest in the L1?
2. Is the development of lexical richness across the first year of study faster in the L2 than in the L1?
3. Does the development of L2 lexical richness benefit from immersion in that L2 environment?

5.2 METHODS

5.2.1 Participants

We obtained data of 675 students who were enrolled in the first year of the Psychology bachelor at Radboud University, Nijmegen (2016-2017). These data included the students' nationality and their preferred language of communication (see Data, 5.2.3), but Radboud University holds no records of students' native language. We therefore worked from the assumption that Dutch was the (only) L1 of students with Dutch nationality, and German was the (only) L1 of students with German nationality. To strengthen this assumption, we excluded 12 students with a double nationality, as well as one Dutch student in the Dutch track who had chosen German as his/her preferred language of communication. In addition, we excluded 44 students whose first nationality was not Dutch or German, because they fell outside the scope of this study. We also excluded four students who did not give us permission to use their data (see Ethics and data handling, 5.2.2). This left 614 students in the data set. For a subset of 362 out of these 614 students, the three exam answers and grade per answer were available. From this subset, we excluded a further 47 students who had written one or more very short answers. The cut-off point for answers being considered too short was set at twenty words, since this seemed the best compromise between still retaining a large enough sample size, and being able to perform a meaningful lexical analysis on the answers. The remaining 315 students formed the participant sample in this chapter. Their descriptives are given in Table 1.

Table 1. Demographic information of the 315 students in the analysis of lexical richness.

Nationality	Track	*n*	% female	Mean age (*SD*)	Age range
Dutch	Dutch	117	85%	19.31 (1.57)	17 – 27
	English	22	64%	19.35 (1.54)	17 – 25
German	Dutch	20	70%	20.83 (1.91)	18 – 26
	English	156	73%	20.39 (1.66)	17 – 27
	Total	315	77%	19.95 (1.73)	17 – 27

Note. 'Age' refers to the students' age on 26 October 2016, the day of the first of the three exams.

5.2.2 Ethics and data handling

Before we obtained the participants' data, we had e-mailed all Dutch and German first-year psychology students at Radboud University to inform them about the current study. In that e-mail, we explained the goal of the study, and which data we were planning to collect. The students were asked to reply to our e-mail in case they did not allow us to collect their data. Four students sent such a reply. The Psychology department then provided the data to us, and we immediately removed all data of the four students who had opted out, as well as the data of the students who did not have either the Dutch or German nationality.

From the remaining data, we removed student numbers and replaced them by anonymised subject codes. The data were protected with a password. Another, password-protected key contained the links between student numbers and subject codes. It was necessary to store the student numbers in order to be able to link the hand-written exams to the demographic information.

The approach described in this section was approved by the ethical commission of the Faculty of Social Sciences of Radboud University (application number ECSW2014-0109-245a). We obtained this ethical approval before starting the data collection.

5.2.3 Data

The data set for Chapter 5 consisted of the following elements:

- Demographic information
  - Student number
  - Gender
  - Date of birth
  - Nationality (one or more)
  - Study language (Dutch/English)
  - Preferred language of communication (Dutch/English/German). Students choose one when registering for a university in the Netherlands – it is not necessarily the same as their native language.
- Hand-written answers to three open exam questions, together with the grades that the course lecturer or teaching assistant had assigned to those answers. This grade reflects the content of an answer, and is not a linguistic score.
  - Exam 1 (course: General Introduction to Psychology, part A; date: 26 October 2016). Question: "Discuss Whorf's language theory. Include the following terms in your answer: Strong and weak variations of the theory."
  - Exam 2 (course: Statistics I, partial exam A; date: 2 February 2017). Question: "Describe a study that is discussed in the book, in which the fact that people are sometimes insensitive to 'sample size' is highlighted. Describe the study's design and the results. (Merely providing an example is insufficient here. By a study we mean something where data have been collected and which has been published.)"

- Exam 3 (course: Statistics I, partial exam C; date: 21 April 2017). Question: "Explain how the 'Good Story Heuristic' may have a negative influence on students' answers in a Statistics I part C exam when they draw up a basic report. Indicate as accurately as possible 1) the section of a basic report that is involved, 2) what are the heuristic changes in the thought process, and 3) the aspect in which their answer will be less good."

Our use of hand-written exams ensured that all students had written their texts under the exact same circumstances (e.g., at the same time and place), and that they did not receive any corrective feedback on their writing. The students in the Dutch and English tracks were taught by the same lecturers. For both courses (Psychology and Statistics I), this was a male Dutch native speaker who spoke English as an L2. The work groups that accompanied the lectures were not necessarily taught by the same teachers in both tracks, and these teachers had various nationalities and L1s.

5.2.3.1 Digitising and correcting the answers to the exam questions

The hand-written answers were entered into the computer by three research assistants (Dutch students at Radboud University) who were paid for this work. One of the research assistants had already finished a bachelor's and master's degree in English language and literature. The other two assistants were bachelor students in psychology. The research assistants created two versions of each answer: one literal transcription, and one corrected transcription. They received the following instructions to correct transcriptions:

- Correct spelling errors, including capitalisation errors. For example, *a Hospital* became *a hospital*.
- Write out abbreviations, so that students who tended to abbreviate more words would not receive different lexical richness scores. For example, *wouldn't* became *would not*.
- Write out numbers within words. For example, *2fold* became *twofold*.
- Remove erroneous duplicate words. For example, *it was was good* became *it was good*.
- Remove erroneous white space. This was mostly relevant in the Dutch data, since compound words should be spelled as one word (in contrast to English). For example, Dutch *moeder taal* (English: *mother tongue*) became *moedertaal*.
- Tag those words that were not written in the study language. For the Dutch exams, the tags *_english* and *_german* were used; for the English exams, *_dutch* and *_german*. This allowed us to later remove such words in order to gauge someone's lexical richness in a particular language (i.e., Dutch or English). Words were not removed if there was no native alternative (e.g., the English loanword *baby* in Dutch these days has replaced the native Dutch *zuigeling*).
- Tag non-existing words, such as *relativates*, with the tag *_nonexisting*.
- Tag illegible words with the tag *_illegible*.

Grammatical errors (e.g., selecting the wrong gender for the Dutch definite determiner) were not corrected, since this did not affect the measures of lexical richness. All corrected transcriptions were seen by two of the three research assistants, in case the first one had failed to detect or adequately correct some mistakes.

5.2.3.2 Tokenising, part-of-speech tagging and lemmatising

The digitised student answers were further preprocessed in Python 3.6 (Python Software Foundation, 2018). Words that had been tagged *_german*, *_english* (in the Dutch data), *_german*, *_dutch* (in the English data), *_nonexisting* or *_illegible* were removed from the data set. The answers were then tokenised (i.e., character sequences were converted into sequences of word tokens), the tokens were tagged with their part of speech (e.g., verb, noun), and lemmatised (e.g., *thinking* became *think*). This was necessary for the subsequent analysis of lexical richness. For the English data, the tokenisation, part-of-speech tagging and lemmatisation were implemented using the Natural Language Toolkit (version 3.2.5; Bird, Loper & Klein, 2009). For the Dutch data, we used Frog (Van den Bosch, Busser, Daelemans & Canisius, 2007).

5.2.3.3 Measures of lexical richness

As explained in the Introduction to this chapter, Lu (2012) listed a total of 26 measures of lexical richness, which together cover the first three lexical richness dimensions that are also used in the present study. While there is only one measure for the dimension of lexical density, there are five for lexical sophistication and twenty for lexical variation. All those measures except one can be automatically calculated with the Lexical Complexity Analyzer (LCA) software (Lu, 2013).

Lexical density is calculated as the ratio between the number of lexical words and the total number of words in a text. The LCA considers nouns, verbs, adjectives and adverbs to be lexical words (with the exception of the verbs *be* and *have*, or their Dutch translations). Since only one measure of lexical density was available (called *Lexical Density*, abbreviated LD), we selected it by default. In Lu's (2012) study, this measure did not correlate significantly with the Mandarin speakers' English L2 oral proficiency scores. However, we do not necessarily consider this a concern, since Mandarin speakers' oral L2 English proficiency seems quite different from Dutch and German students' writing proficiency in Dutch and English. Furthermore, we had the first research question acting as a methodological check of our selected lexical richness measures.

Lexical sophistication is calculated as the ratio between the number of 'sophisticated' words and the total number of words in a text. 'Sophisticated' in itself is a subjective term and can be defined in various ways (see Lu, 2012, for an overview), but in the LCA words are considered sophisticated if they do not appear in the top-2000 most frequent English words (as computed from the British National Corpus; the relevant frequency table was included

in the LCA software distribution). For the Dutch data, we used the CELEX lexical database (Baayen, Piepenbrock & Gulikers, 1995) to construct a frequency table of lemmas.

Of the five available sophistication measures, three pertained to verb sophistication, and the remaining two to the lexicon as a whole. Although the three verb-related measures all significantly correlated with the Mandarin speakers' English proficiency in Lu (2012), and the two other measures did not, we still opted for one of the latter two measures since our interest was not restricted to just verbs. Of these two measures, one is computed as the ratio of sophisticated word tokens to all word tokens, and one is computed as the ratio of sophisticated word types to all word types. Word types are all the unique words in a text, whereas word tokens encompass all word occurrences, including multiple repetitions of the same word. We selected the measure based on word types, called *Lexical Sophistication-II* (LS2), because whether someone uses a sophisticated word once or many times does not seem to be as informative of his/her degree of lexical sophistication.

Lexical variation can be calculated in various ways, all reflecting the diversity of the vocabulary that is used. The first approach focuses on the absolute number of different words that is used in a text. Different methods to correct for text length are available, for example considering samples of a fixed length only. The second approach consists of various measures of type-token ratio. Again, there are different options to correct for text length (longer texts tend to have lower type-token ratios). Finally, there is a range of measures that concern the ratio between part-of-speech tokens and part-of-speech types.

Again, we wanted to select a measure that concerned all words rather than certain parts of speech only. This ruled out all the measures based on specific parts of speech. Out of the remaining measures, we selected *Number of Different Words (Expected Sequence)* (NDW-ES). This measure is calculated as the mean number of word types in a large number of randomly drawn word sequences of a fixed length. We used 10,000 samples of twenty words; they were automatically drawn with replacement from each student's exam answer. We chose twenty as the sample length because our shortest samples consisted of twenty words (see the Participants section, 5.2.1); it also seemed long enough to be representative. As compared to the other available measures of lexical variation, the NDW-ES measure seemed most robust because of the resampling. It also correlated significantly with the Mandarin speakers' English proficiency ($\rho = .32$, $p < .001$) in Lu (2012).

All three selected measures are by definition independent of text length. This was an important criterion, because the mean length of the written answers varied per exam (see Table 2 in the Results section) and per student. To be completely sure, we also calculated the correlations between answer length and our lexical richness scores, and found that they were low and non-significant.

5.2.4 Analysis

5.2.4.1 Design

There were three dependent variables, namely the three measures of lexical richness as presented above. The independent variables were Exam, Group and Grade (we will write variable names with a capital letter). Exam was a within-participants variable indicating the exam (and thereby, moment in time) from which a writing sample originated. It had three levels: Exam 1 (October), Exam 2 (February), and Exam 3 (April). Group was a between-participants variable with four levels: Dutch in Dutch track, Dutch in English track, German in Dutch track, and German in English track. We preferred such a Group variable over the alternative of using a two-by-two design with the variables of Nationality (Dutch, German) and Study language (Dutch, English). The reason for this is that the main effects of Nationality and Study language could not be meaningfully analysed due to the unequal samples sizes in the four groups (see Table 1): Nationality would be a confounding variable in the interpretation of Track, and vice versa. Using the Group variable with four levels allowed us to more effectively make use of planned contrasts, which are explained in section 5.2.4.4. Grade was a within-participants variable. It is the grade that was assigned by the course lecturer to the content of a student's answer on an exam question (not the linguistic quality or linguistic accuracy of the answer). It was included as a predictor because some questions seem to have been more difficult than others (see Table 2 in the Results section), and we wanted to statistically account for a possible relationship between Grade and lexical richness (see section 5.2.4.3) before evaluating the effect of Group and its interaction with Exam.

5.2.4.2 Controlling Type-I error rates

We started out with $\alpha = .05$. Following Lakens (2016), we controlled Type-I error rates within each research question, but not between the questions. Each question was evaluated on each of the three dependent variables (i.e., lexical density, sophistication and variation). Thus, we could have used a Bonferroni correction and divided .05 by 3, yielding $\alpha = .0167$. However, the Bonferroni correction is quite conservative because it does not take into account the potential correlation between the variables on which the tests are performed. In our case, the correlation between lexical density and lexical sophistication was $r = .38$, $p < .001$. The correlation between lexical density and lexical variation was $r = .06$, $p = .33$, and the correlation between lexical sophistication and lexical variation was $r = .01$, $p = .86$. These correlations were calculated by averaging the lexical richness values over the three exams.

Alternatively, Nyholt (2004) provides software that calculates the significance threshold required to keep Type-I error rates at 5%, while taking the correlation between the dependent variables into account. Li and Ji (2005) made further improvements to Nyholt's method, and Nyholt (2015) recommends that Li & Ji's estimate is used if it is smaller than Nyholt's estimate. In our case, Nyholt's significance threshold was .0173 and Li and Ji's threshold was .0170. Thus, we set α to .0170.

5.2.4.3 Model comparisons

We used linear-mixed effects models with Exam, Group and Grade as fixed effects. For each of our three measures, we created the following models:

1. lexical richness measure ~ 1 + Exam + (1 | Subject)
2. lexical richness measure ~ 1 + Exam + Grade + (1 | Subject)
3. lexical richness measure ~ 1 + Exam (+ Grade) + Group + (1 | Subject)
4. lexical richness measure ~ 1 + Exam (+ Grade) + Group + Exam : Group + (1 | Subject)

In this notation, '~' indicates that the lexical richness measure is modelled from the terms that follow. '1' represents the intercept, the terms following the 1 represent the fixed effects with ':' representing an interaction effect, and '(1 | Subject)' represents random intercepts at the subject level. We included these random intercepts so that individual variation between the students became part of the model rather than ending up in the error term. It was not possible to also include a random effect of Exam at the subject level (i.e., Exam | Subject), because Exam was a categorical variable and therefore the number of estimated random effects would equal the total number of observations (945, namely 3*315). This would be problematic, because the estimated random effects would be confounded with the residual variation.

We investigated the significance of the fixed effects through model comparisons with likelihood ratio tests. The effect of Grade was investigated by comparing the fit of Model 1 and Model 2 to the data. If the second model was found to be significantly better, we kept Grade in the model. If not, if was removed, because it was not of interest to our research questions. To investigate the main effect of Group, the fit of Model 3 to the data was compared to that of either Model 1 or 2 (depending on whether Grade was included or not). The significance of the interaction between Exam and Group was examined by comparing Models 3 and 4. All statistics were carried out in R (R Core Team, 2018).

5.2.4.4 Planned contrasts and pairwise comparisons

If Model 3 was found to fit the data significantly better than the previous model, the model estimates were subjected to planned contrasts in order to answer Question 1. We compared the lexical richness scores of the Dutch students in the Dutch track to that of the three other groups.

If the comparison between Model 3 and 4 was significant, we performed pairwise comparisons on Model 4 in order to answer Questions 2 and 3. We opted for pairwise comparisons rather than planned contrasts so that all possible contrasts could be investigated (i.e., Exams 1 versus 2, 1 versus 3, and 2 versus 3). This means that an additional correction for the inflation of Type-I error rates was needed. To this end, we used the Benjamini-Hochberg (1995) procedure, which controls the false discovery rate. This procedure involves ranking all the *p*-values that result from the pairwise comparisons from the smallest to the largest.

Let *k* be the total number of comparisons made, and *j* the rank of the *p*-value (where the smallest *p*-value has rank 1, and the largest has rank *k*). For each *p*-value, α was set to j/k^* .0170 (α had been set to .0170 to begin with as there were three dependent variables, see 5.2.4.2).

5.2.4.5 Model diagnostics

On all measures, we examined whether the assumptions of linear regression had been met for Models 3 and 4, because these were the models we performed planned contrasts and pairwise comparisons on. This involved plotting the model's predicted values by the model's residual values. This plot should not show any obvious patterns. We also inspected whether the assumption of homoscedasticity was met (the standard deviations of the residuals should not depend on the x-value of the predicted values). Next, we inspected whether the residuals were normally distributed, although Winter (2013) points out that this assumption is the "least important" (p. 18) and is not even mentioned by Gellman and Hill (2007). We investigated the presence/absence of influential data points by calculating Cook's distance (Cook, 1977). Following McDonald (2002), we considered values >0.85 as a reason for concern. The independence assumption of linear models was taken care of by the random-subject intercepts in all models. As for the absence of collinearity, our predictors of interest (Exam and Group) were always independent because our design was fully balanced (i.e., all students in the data set took part in all exams). Finally, we investigated whether the random-subject intercepts were normally distributed. The outcomes of this process are reported in Appendix A. Generally speaking, all model diagnostics looked good, with the exception of Model 3 for lexical variation. However, we are not very concerned about that particular model because we did not use it for any planned contrasts (see 5.3.4.2).

5.3 RESULTS

5.3.1 Exam descriptives

Table 2 shows the descriptive statistics for the writing samples from each exam. It can be seen that the third exam yielded shorter answers that were graded substantially higher. But since the grades are included in our statistical models (see section 5.2.4.1), this is not problematic.

Table 2. Descriptive statistics for the students' answers on the three open exam questions.

Exam	Mean text length in words (*SD*)	Range text length	Mean grade (*SD*)
1 (October)	114 (60)	20 – 377	7.10 (2.74)
2 (February)	100 (28)	32 – 197	7.49 (3.79)
3 (April)	59 (18)	20 – 121	8.35 (3.02)

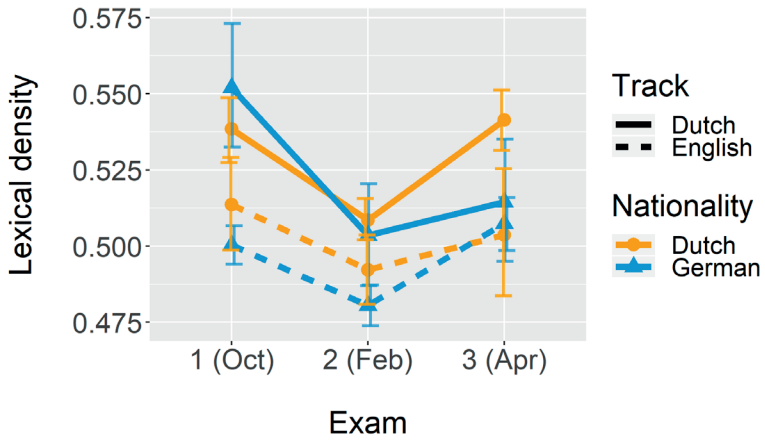
Note. In all cases, the range of the grades was 0-10. Text length was calculated after preprocessing the answers. Standard deviations were obtained through bootstrapping with 10,000 iterations[1].

[1] The concept of bootstrapping is explained in detail in Chapter 6, sections 6.2.4.2 and 6.2.4.3.

5.3.2 Lexical density

Figure 1 shows the lexical density scores as measured from the three exams. The overall pattern in all four groups is striking: Rather than there being a positive development over time, the lexical density scores decrease between Exam 1 and 2, and then rise again. A quick glance ahead to Figures 2 and 3 reveals similar patterns for the two other measures of lexical richness. This finding could likely be explained by the three exam questions not being comparable in their topic and complexity, which will be considered in more detail in the Discussion (5.4.4). Therefore, and as planned, we will only focus on the main effect of Group, and its interaction with Exam.

Figure 1. Lexical density across the three exams.



5.3.2.1 Model comparisons

Model comparisons showed no significant effect of Grade ($\chi^2 = 0.82$, $df = 1$, $p = .36$), but a main effect of Group ($\chi^2 = 73.15$, $df = 3$, $p < .001$). The interaction between Exam and Group was non-significant ($\chi^2 = 12.13$, $df = 6$, $p = .059$), with α being .0170 (see section 5.2.4.2).

5.3.2.2 Is lexical density the highest in the L1?

Lexical density was significantly higher for the Dutch students in the Dutch track as compared to the Dutch students in the English track ($b = 0.026$, $SE = 0.007$, $t = 3.63$, $p = .0003$), and also as compared to the German students in the English track ($b = 0.033$, $SE = 0.004$, $t = 8.76$, $p < .001$). There was no significant difference between the Dutch and German students in the Dutch track ($b = 0.006$, $SE = 0.008$, $t = 0.81$, $p = .42$).

5.3.2.3 Is the development of lexical density slower in the L1 as compared to the L2?

The interaction between Exam and Group was not significant. Thus, we found no evidence for a differential development of study language lexical density between the Dutch students in the Dutch and English tracks.

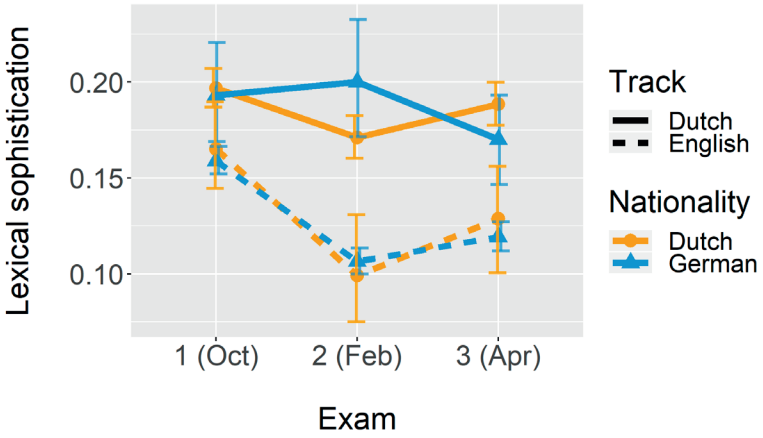
5.3.2.4 Does the development of L2 lexical density benefit from immersion in the L2?

Because the interaction between Exam and Group was not significant, we also did not further examine the development of L2 lexical density between the German students in the Dutch and English track (i.e., German students who are and who are not exposed to the L2 outside of the university context). Thus, there was no evidence for a differential development of L2 lexical density as a function of L2 immersion.

5.3.3 Lexical sophistication

Figure 2 shows the lexical sophistication scores as measured from the three exams. For all groups except the German students in the Dutch track, the scores at Exam 2 are lower than those at Exams 1 and 3. The score of the German students in the Dutch track, on the contrary, rises from Exam 1 to 2. However, the wide confidence intervals indicate that the confidence regarding the precision of this estimate is relatively low; please recall that the sample size for this group was only 20 students. A clear distinction is visible between the scores of the students who study in Dutch versus English (seemingly regardless of nationality/L1).

Figure 2. Lexical sophistication across the three exams.



5.3.3.1 Model comparisons

There was a significant main effect of Grade (χ^2 = 14.25, *df* = 1, *p* < .001). Thus, we kept Grade in the model. In addition, the main effect of Group was significant (χ^2 = 159.69, *df* = 3, *p* < .001), as well as the interaction between Exam and Group (χ^2 = 35.64, *df* = 6, *p* < .001).

5.3.3.2 Is lexical sophistication the highest in the L1?

Lexical sophistication was significantly higher for the Dutch students in the Dutch track as compared to Dutch students in the English track (*b* = 0.055, *SE* = 0.008, *t* = 6.81, *p* < .001), and also as compared to the German students in the English track (*b* = 0.056, *SE* = 0.004, *t* = 13.19,

$p < .001$). There was no significant difference between the Dutch and German students in the Dutch track ($b = -0.002$, $SE = 0.008$, $t = -0.27$, $p = .79$).

5.3.3.3 Is the development of lexical sophistication slower in the L1 as compared to the L2?

As can be seen from Table 3, on none of the exam contrasts there was a significant difference between the development of lexical sophistication of the Dutch students in the Dutch track and the Dutch students in the English track.

Table 3. Pairwise comparisons for the development of lexical sophistication (i.e., changes in lexical sophistication from one exam to the next) between the Dutch students in the Dutch track and the Dutch students in the English track.

Exam comparison	Mean difference (*SE*)	Test statistics	Corrected α
1 versus 2	-0.040 (0.016)	$t = -2.52$, $p = .012$	1/3 * .0170 = .0057
2 versus 3	0.009 (0.016)	$t = 0.54$, $p = .59$	3/3 * .0170 = .0170
1 versus 3	-0.032 (0.016)	$t = -1.98$, $p = .048$	2/3 * .0170 = .0113

Note. See section 5.2.4.4 for an explanation of the α-correction.

5.3.3.4 Does the development of L2 lexical sophistication benefit from immersion in the L2?

The lexical sophistication scores of the German students in the Dutch track slightly increased between Exam 1 and 2, while the scores of the German students in the English track decreased. This interaction was significant (see Table 4) and in the expected direction. However, the direction of the interaction then reversed, while the interaction remained significant. Between Exam 2 and 3, the scores of the German students in the English track slightly increased, while the scores of the German students in the Dutch track decreased (see Table 4).

Table 4. Pairwise comparisons for the development of lexical sophistication between the Dutch and German students in the Dutch track.

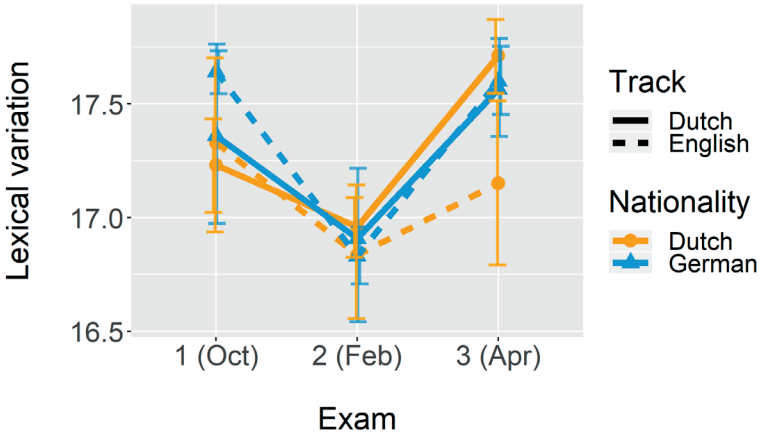
Exam comparison	Mean difference (*SE*)	Test statistics	Corrected α
1 versus 2	-0.063 (0.017)	$t = -3.81$, p **< .001**	1/3 * .0170 = .0057
2 versus 3	0.045 (0.017)	$t = 2.69$, $p =$ **.007**	2/3 * .0170 = .0113
1 versus 3	-0.019 (0.017)	$t = -1.12$, $p = .26$	3/3 * .0170 = .0170

Note. Significant *p*-values are printed in bold.

5.3.4 Lexical variation

Figure 3 shows the lexical variation scores as measured from the three exams. This time, no difference between the four groups is immediately apparent, although the Dutch students in the English track seem to score lower on lexical variation in Exam 3. The inferential statistics presented below will tell us whether this difference based on visual inspection also reached statistical significance.

Figure 3. Lexical variation across the three exams.



5.3.4.1 Model comparisons

Model comparisons showed no significant effect of Grade (χ^2 = 2.51, *df* = 1, *p* = .11), and neither a main effect of Group (χ^2 = 4.80, *df* = 3, *p* = .18). The interaction between Exam and Group was significant (χ^2 = 21.61, *df* = 6, *p* = .0014).

5.3.4.2 Is lexical variation the highest in the L1?

Since there was no main effect of Group, we did not perform any further analyses. There is no evidence that lexical variation scores are higher for L1 as compared to L2 speakers.

5.3.4.3 Is the development of lexical variation slower in the L1 as compared to the L2?

As can be seen from Table 5, there was no significant difference in the development of lexical variation between the Dutch students in the Dutch track and the Dutch students in the English track.

Table 5. Pairwise comparisons for the development of lexical variation (i.e., changes in lexical variation from one exam to the next) between the Dutch students in the Dutch track and the Dutch students in the English track.

Exam comparison	Mean difference (*SE*)	Test statistics	Corrected α
1 versus 2	-0.22 (0.27)	*t* = -0.81, *p* = .42	3/3 * .0170 = .0170
2 versus 3	-0.44 (0.27)	*t* = -1.60, *p* = .11	2/3 * .0170 = .0113
1 versus 3	-0.66 (0.27)	*t* = -2.42, *p* = .016	1/3 * .0170 = .0057

Note. See section 5.2.4.4 for an explanation of the α-correction.

5.3.4.4 Does the development of L2 lexical variation benefit from immersion in the L2?

There was no significant difference in the development of lexical variation between the German students in the Dutch track and the German students in the English track (see Table 6).

Table 6. Pairwise comparisons for the development of lexical variation between the Dutch and German students in the Dutch track.

Exam comparison	Mean difference (*SE*)	Test statistics	Corrected α
1 versus 2	-0.36 (0.28)	t = -1.29, p = .20	1/3 * .0170 = .0057
2 versus 3	0.11 (0.28)	t = 0.41, p = .68	3/3 * .0170 = .0170
1 versus 3	-0.25 (0.28)	t = -0.88, p = .38	2/3 * .0170 = .0113

5.4 DISCUSSION

5.4.1 Is lexical richness the highest in the L1?

Generally speaking, lexical richness was the highest in the L1: The Dutch students in the Dutch track obtained higher scores on lexical density and lexical sophistication than both the Dutch and the German students in the English track. However, in two aspects the results departed from this outcome. To begin with, on lexical variation no significant difference between the Dutch students in the Dutch track and any of the other three groups could be detected. In addition, the Dutch and German students in the Dutch track did not differ significantly on any of the measures.

The first research question had been intended as a methodological check, and we had expected to find that the Dutch students in the Dutch track would outperform the other three groups on all measures. As we did not always find this to be the case, we will consider some possible causes for the null effects, beginning with the fact that we found no difference between Dutch and German students in the Dutch track on any of the measures.

5.4.1.1 No difference between Dutch and German students in the Dutch track

If we assume that the German students indeed were L2 speakers of Dutch, and therefore must have been less proficient than the L1 Dutch native speakers, then the lexical richness measures apparently were not sensitive enough to detect the differences between these two groups. Perhaps the non-native status of the German L2 speakers of Dutch would be more visible in their lexical or grammatical errors, or aspects of language proficiency other than lexical richness.

Alternatively, we can question the assumption that Dutch was an L2 to the German students. Perhaps some of them had a (near-)native command of Dutch to begin with, making them practically indistinguishable from the students with the Dutch nationality in terms of Dutch proficiency. Radboud University attracts many German students from the Dutch-German border region, where there are relatively many Dutch-German families. Good or (near-)native Dutch language skills could also explain the relatively atypical choice, made

by only 11% of the German students in our sample, to enrol in the Dutch rather than the English track.

5.4.1.2 No effect on lexical variation

Concerning lexical variation, the Dutch students in the Dutch track did not differ from any of the other three groups. Apparently, the native speakers in our sample did not use more different words than the non-native speakers. This is opposite to the outcome of Lu (2012), who presented a significant correlation between oral L2 proficiency and this specific measure of lexical variation (i.e., NDW-ES). No obvious explanation for this finding comes to mind, although one should remember that our study differed from Lu's in several aspects. He looked for *within-group* correlations between scores on lexical richness measures and ratings of L2 English oral proficiency of Mandarin native speakers, while we looked for *group-level* differences in lexical richness scores calculated from written texts, in Dutch and German speakers. Any of these differences in study design, target language and participant population could underlie our finding.

5.4.1.3 A main effect of language?

Despite the above exceptions, for the most part our hypothesis of the Dutch students in the Dutch track scoring higher on the lexical richness measures than the other groups was supported by our data. In other words, an L1 versus L2 effect seemed visible. However, there is also the alternative explanation that we may not have detected an effect of nativeness, but rather a main effect of language (lexical richness in Dutch versus English). Although English and Dutch both are West-Germanic languages, there are some typological differences between them that could have an effect on lexical richness scores. For example, compound nouns are written as two words in English (e.g., *mother tongue*), but as one word in Dutch (e.g., *moedertaal*). This could lead to higher lexical density scores in English, because there are relatively more separate nouns in a text, and thus a higher proportion of content words. Lexical sophistication scores could also be affected. Since compounds usually are less frequent than their constituents, they are more likely to count as sophisticated. However, the difference in compound spelling between English and Dutch is just one of the many differences between the two languages that could affect lexical richness scores. To our knowledge, these differences have not been mapped out to the extent that systematic predictions could be made about whether lexical richness scores by default should be higher in Dutch or in English.

Something else to note is that during preprocessing we had removed English and German words from the Dutch texts, and Dutch and German words from the English texts. This was done because we were exclusively interested in the students' lexical richness in their study language, and not in their lexical knowledge of other languages. Using an English loan word rather than its Dutch counterpart (in a Dutch text) could even indicate that someone did not know the word in Dutch and therefore had to resort to English. We removed 110

English words from Dutch texts, but only one Dutch word from an English text (from both Dutch and English texts, one German word was removed). Since the vast majority of the removed English words were content words, we can wonder whether this could have caused a decrease in lexical density scores on Dutch texts.

We ran a computer simulation to investigate this option. It showed that the removal of English words from Dutch texts only had a very small impact on lexical richness scores. In total there were 411 Dutch texts, with an average length of 77 words. The fact that we removed a total of 110 English words means that only 1 in about 288 words was removed from the Dutch texts. If we make the conservative assumption that all of the removed words were content words, Dutch lexical density scores would have decreased by 0.33% (and in an absolute sense, from an average of ±0.512 to ±0.510). Also assuming that all of the removed words were unique in their 20-word window, lexical variation scores would have decreased by 0.05% (in an absolute sense, from an average of ±17.35 to ±17.34).[2] Thus, the effect would have been minimal. We did not inspect lexical sophistication because there was no reason to assume that either sophisticated or unsophisticated words would have been overrepresented in word removal. But even if one of the two types was in fact overrepresented, the magnitude of the effect would be at most the size of that of lexical density.

5.4.1.4 Conclusions of the methodological check

Regarding the use of the lexical richness measures for answering our second and third research question, we draw the following conclusions. First, the finding that the lexical variation measures could not distinguish Dutch L1 native speakers from any of the other L2 groups suggests that this measure was less suitable for our purposes. While we will still include lexical variation in the rest of the Discussion, one should keep its potential limitations in mind. Second, we no longer assume that the German students in the Dutch track necessarily spoke Dutch as an L2. This has some implications for the interpretation of the data regarding Question 3.

Any potential main effects of language on lexical richness do not concern us very much with regard to Questions 2 and 3, because in neither of these questions we look at main effects. More specifically, we do not simply consider whether the scores are higher for students in the Dutch or in the English track (which could be problematic), but we look at whether the scores develop differently between the four groups. Even if there were a main effect of language on lexical richness, this should not, or hardly, affect lexical richness growth over time.

[2] The file containing the code to replicate this simulation is called simulation_no_removal.R and can be found at https://github.com/johannadevos/StudyLanguage.

5.4.2 Is the development of lexical richness slower in the L1 as compared to the L2?

On none of the three measures we found evidence for differential development in study language proficiency between Dutch students in the Dutch and English tracks. Thus, studying in English does not lead to a quicker (or slower) development of lexical proficiency than studying in Dutch. This means that Dutch students who are hoping to also improve their language skills whilst studying psychology (or another degree that is offered in both Dutch and English) can gain as many Dutch language skills from studying in Dutch, as English language skills from studying in English. It may therefore be reasonable for them to base their choice of study language on the job market they want to prepare themselves for. In addition, in Chapter 6 we will see that there seems to be an advantage to studying in the L1 when it comes to the grades that students obtain.

5.4.3 Does the development of L2 lexical richness benefit from immersion in the L2?

For the German students, we had expected to see a superior development of Dutch as compared to English lexical richness, thanks to the students being immersed in a Dutch language environment in Nijmegen. However, for the most part our hypothesis could not be confirmed: We found no significant effects on lexical density and variation. On lexical sophistication, two out of three contrasts were significant, but these two effects were in opposite directions.

5.4.3.1 Opposite effects on lexical sophistication

Between Exam 1 and 2, the German students' lexical sophistication scores dropped in the English track, but not in the Dutch track. The direction of this interaction effect was in line with our hypothesis, with an advantage for the German students in the Dutch track. However, between Exam 2 and 3 the direction of the effect reversed, indicating a stronger development of lexical sophistication in English than in Dutch. Both these effects depend on a seemingly outlying data point, namely a very high lexical sophistication score of the German students in the Dutch track at Exam 2. It is the only data point in the entire data set (including all other groups and lexical richness measures) where the score on Exam 2 was higher than on Exam 1. The large confidence intervals around this data point (due to the low number of German students in the Dutch track) indicate the high uncertainty that is associated with this estimate. In order to get more insight in the reliability of the outlying data point, we inspected the distribution of the scores. We found that the German students in the Dutch track did not obtain higher scores than Dutch students in the Dutch track. Rather, what caused their average score to be higher, was the fact that they obtained relatively fewer scores in the lower segment (roughly, 0-0.15). Thus, it was not the case that the outlying data point was caused by a few German students scoring excessively high. All in all, then, no immediate explanation for the reversal of the effect comes forward.

5.4.3.2 No effect on lexical density and variation

On the measures of lexical density and variation, we detected no significant difference between the German students in both tracks. As a potential explanation, we note that some German students who study in Nijmegen live in Germany rather than in the Netherlands (since Nijmegen is close to the German border). In addition, even the German students who are living in Nijmegen may have German or other international house mates and friends, meaning that they would hear and speak little Dutch outside the university. It is also conceivable that in terms of music, films, books and video games, German students encounter more English than Dutch.

In our other studies (Chapters 3 and 4), all native German participants filled out a survey on their language background, among other things providing information on their living situation. Of those 124 participants, about 87% lived in the Netherlands, 66% had Dutch house mates and 73% had Dutch friends. While this indicates that is seems reasonable to expect that at least the majority of the German students had some ties to the Netherlands in addition to being a student at Radboud University, it also shows that we cannot necessarily expect this to apply to all students. The data in Chapters 3 and 4 also stem from the period when Psychology was not yet offered in English, when there were relatively fewer German students in Nijmegen, and when learning Dutch was obligatory prior to the first study year.

With regard to the current study, we conclude that the assumption that (a majority of) the German students were immersed in the Dutch language seems reasonable enough, but for future studies it would be better if such information could be confirmed explicitly. A possible explanation for the null effect then might be that even if the German students were immersed in a Dutch language environment, they were also often exposed to English through travelling, international friends and (online) media.

There is one more potential explanation for the null effect. Please recall that even in the Dutch track some of the reading materials were in English. This means that within the study context, the students in the English track had more exposure to English than the students in the Dutch track had exposure to Dutch. Perhaps this surplus of exposure to the study language (English) within the university context has countered the possible effects of immersion in the study language (Dutch) outside of the university context.

5.4.4 Suggestions for future research

It currently remains unknown what the lexical development of first-year students looks like in an absolute sense, because it was clear from the outset that the three exam questions might not have been comparable in topic and complexity. The first exam question concerned Whorf's language theory, while the second and third answer concerned statistical theory. Several studies have shown that there are relationships between the characteristics of the writing task that a learner engages in, and the resulting written text. For example, Frear and Bitchener (2015) found that task complexity was related to the written lexical complexity of intermediate learners of L2 English. Similarly, Yang, Lu and Cushing Weigle (2015) found

a significant relationship between writing topics and syntactic complexity in L2 English writing. As a result, in the current study it was not possible to simply consider, for each group separately, the development of lexical richness scores over time.

It is recommended that in future studies the topic and complexity of the writing samples be more comparable. This would allow researchers to investigate how much (and even whether) students' lexical richness increases over time as a function of naturalistic exposure to the study language. If such research is done, it would be important to also include control groups of students who produce writing samples in the language in which they do not study. For example, obtaining English writing samples of students who study in Dutch would allow the distinction between the effects of studying in English, and the effects of exposure to English outside of the university context.

5.5 SUMMARY AND CONCLUSIONS

In this chapter we presented the first empirical study on the development of lexical richness in first-year university students in the Netherlands. The study showed that the lexical richness of Dutch students is higher in Dutch than in English when it comes to lexical density and lexical sophistication, but not lexical variation. With regard to development over time, we found no evidence that lexical richness developed differently in Dutch students who studied in Dutch versus in English. Similarly, German students studying in Dutch versus English also did not seem to develop their lexical richness more quickly in one language or the other. We have drawn no conclusions about the development of lexical richness in an absolute sense (e.g., does it increase over time, and if so, how much?). In order to be able to do so, it should be a high priority in future research to control the topic and complexity of the students' writing samples, and include control groups of students who provide writing samples in the language that they do not study in.

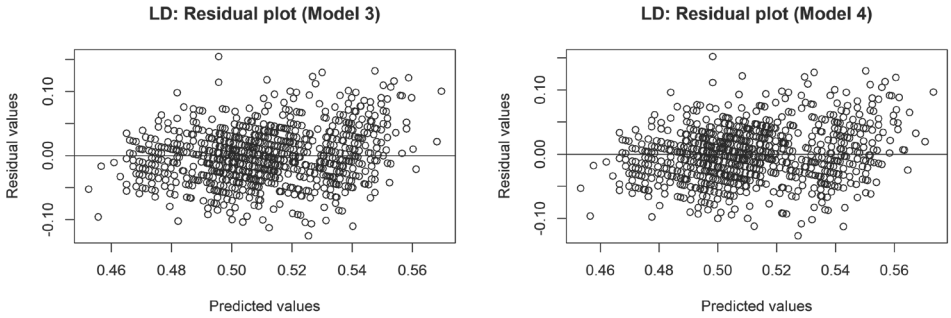
As mentioned throughout the Discussion, our hypotheses often depended on assumptions about the students' linguistic background and (lexical) proficiency in the languages under investigation. If we are to re-evaluate the current hypotheses in future research, it is necessary to collect such background information about the students. For example, the LexTALE lexical decision test (Lemhöfer & Broersma, 2012) is available in Dutch, German and English and takes five minutes to complete. With this information, dual-language programmes such as Psychology at Radboud University, where the exact same programme is offered in two languages, have enormous potential for research on the effects of EMI.

APPENDIX A: MODEL DIAGNOSTICS

Lexical density

Residual plots for Model 3 and 4 are given in Figure A. They look good: The residuals are symmetrically clustered around the y = 0 line, and they are no clear patterns (such as a U-shape). This indicates that there seems to have been a linear relationship between the independent and the dependent values. In addition, the standard deviation of the residuals does not seem to depend on the predicted values, meaning the assumption of homoscedasticity also was met.

Figure A. Residual versus predicted values for lexical density, as predicted from Exam and Group (left), plus their interaction (right).



The residuals also seemed to be normally distributed (see Figure B), and the subject intercepts as well (see Figure C).

Figure B. Histograms of the residuals for lexical density.

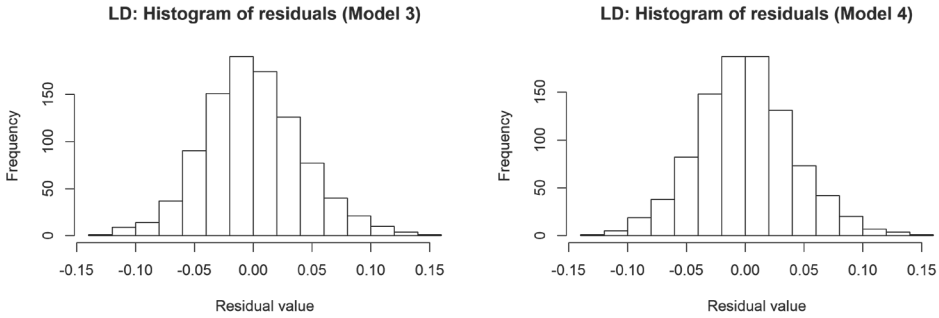
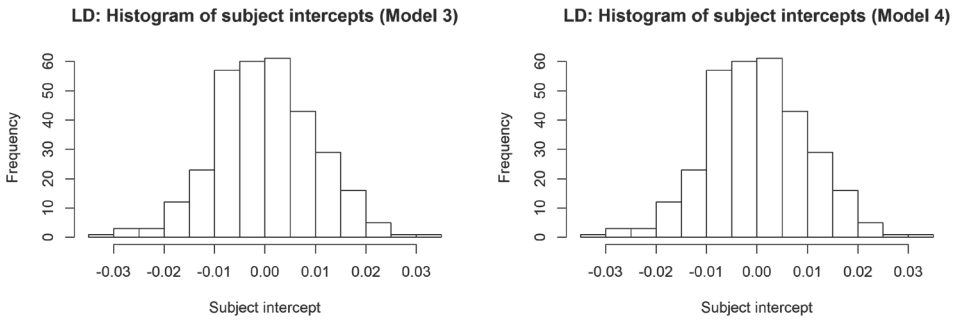


Figure C. Histograms of the subject intercepts for lexical density.

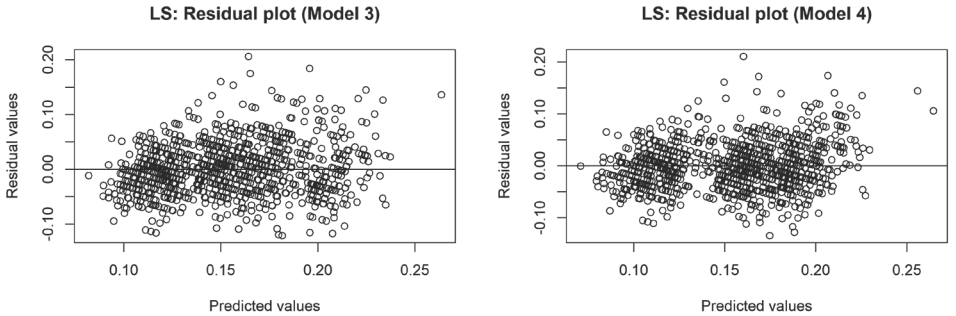


There were no overly influential data points: The highest Cook's distance value for Model 3 was 0.03, and for Model 4 it was 0.02. These values are well-below our cut-off point of 0.85 (see 5.2.4.5). In short, all diagnostics for lexical density looked good, and we considered the models' assumptions to be met.

Lexical sophistication

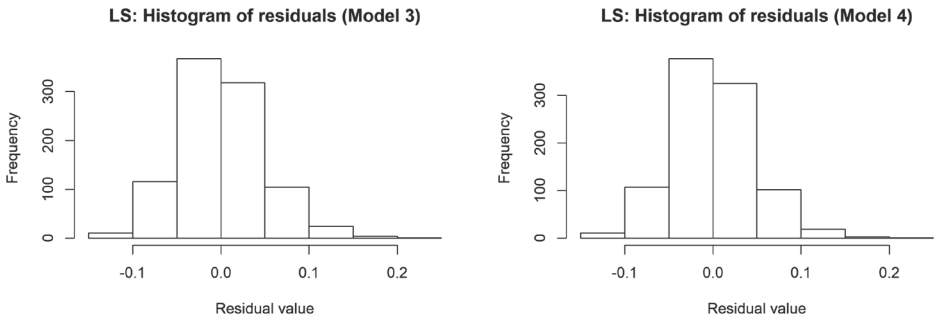
The residual plots for lexical sophistication showed no obvious patterns in the residuals (see Figure D).

Figure D. Residual versus predicted values for lexical sophistication, as predicted from Exam, Grade and Group (left), plus the interaction between Exam and Group (right).



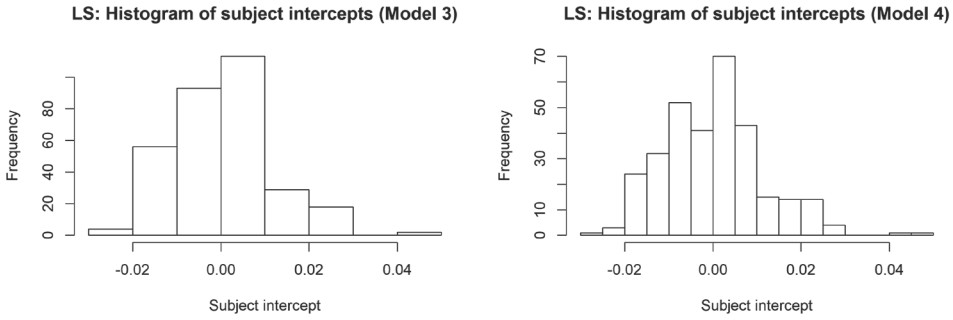
The residuals were not perfectly normally distributed (see Figure E), but the distribution seemed relatively normal. According to Winter (2013, p. 19), it suffices when there are "no obvious violations of the normality assumption", which is the case in Figure E.

Figure E. Histograms of the residuals for lexical sophistication.



The distribution of subject intercepts also looked quite normal. The highest Cook's distance value for Model 3 was 0.18 and for Model 4 it was 0.11; both below the cut-off point of 0.85. Thus, it seems that no data point exerted a disproportionate influence on the model.

Figure F. Histograms of the subject intercepts for lexical sophistication.

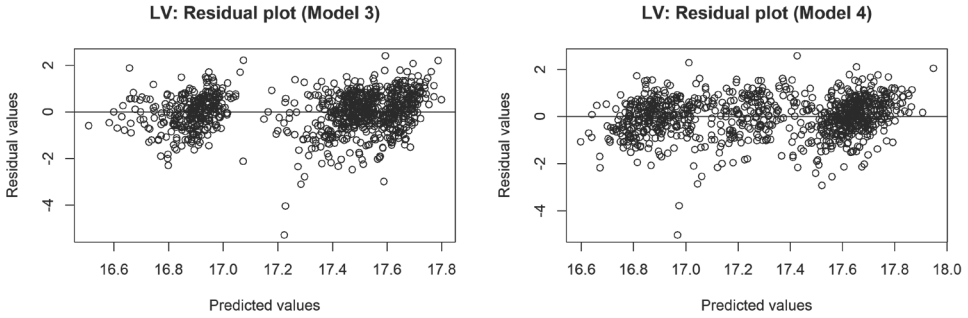


All in all, our models for lexical sophistication seemed reliable.

Lexical variation

Model 3's residual plot showed an unexpected, bimodal pattern (see Figure G); the model hardly predicted any outcomes between (roughly) 17.0 and 17.2. It is unclear what caused this pattern. However, the bimodal pattern largely disappeared after we added the interaction between Exam and Group to Model 3, thereby creating Model 4. We only performed pairwise comparisons on Model 4, but no planned contrasts on Model 3. This was because Model 3 did not fit the data significantly better than its simpler precursor (Model 1). Because we did not use Model 3 for any planned contrasts, the deviations do not concern us very much. In Figure G, as compared to the residual plots for lexical density (Figure A) and lexical sophistication (Figure D), the magnitude of the residuals was much bigger. However, this is simply because lexical variation was measured on a 0-20 scale, whereas both lexical density and lexical sophistication were measured on a 0-1 scale.

Figure G. Residual versus predicted values for lexical variation, as predicted from Exam and Group (left), plus their interaction (right).



For both models, the histograms of the residuals were skewed to the left (see Figure H), but this was due to only a couple of data points. Apart from this, the distribution of the residuals seemed normal.

Figure H. Histograms of the residuals for lexical variation.

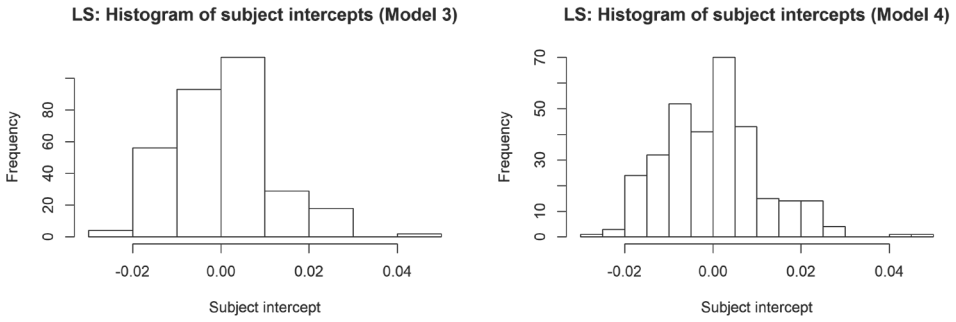
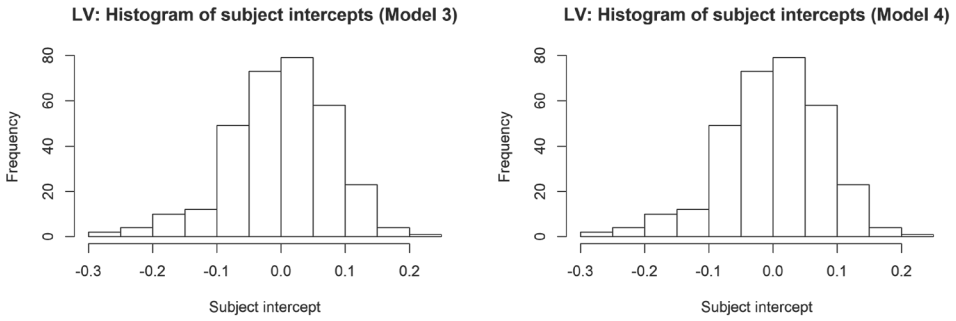


Figure I shows that the subject intercepts were approximately normally distributed.

Figure I. Histograms of the subject intercepts for lexical variation.



No data points exerted a disproportionate influence on the regression line: The highest Cook's distance value for Model 3 was 0.06, and for Model 4 it was 0.04. All in all, the reliability of the lexical variation analysis seemed acceptable.

6.

Does study language (Dutch versus English) influence study success of Dutch and German students in the Netherlands?

This chapter is based on:

De Vos, J. F., Schriefers, H., & Lemhöfer, K. (2019). *Does study language (Dutch versus English) influence study success of Dutch and German students in the Netherlands?* Manuscript submitted for publication.

ABSTRACT

We investigated whether the language of instruction (Dutch or English) influences the study success of 614 Dutch and German first-year psychology students in the Netherlands. The Dutch students who were instructed in Dutch studied in their native language (L1), the other students in a second language (L2). These were Dutch students studying in English, German students studying in Dutch, and German students studying in English. In addition, only the Dutch students (regardless of instruction language) studied in their home country, while the German students studied abroad. Both these variables could potentially influence study success, which we operationalised as the number of European Credits (ECs) the students obtained, their grades, and their drop-out rates. We found that the L1 group outperformed the three L2 groups with respect to grades, but there were no significant differences in ECs and drop-out rates (although descriptively, the L1 group still performed best). Furthermore, we found that the Dutch students who chose to study in L2 English already had better English skills in high school than the Dutch students who chose to study in L1 Dutch, and we controlled for this pre-existing group difference in our analyses. Finally, the students' lexical richness (i.e., productive vocabulary knowledge) did not predict their grades, but one lexical richness measure did predict their drop-out rates. In conclusion, this study showed an advantage of studying in the L1 when it comes to grades, and thereby contributes to the current debate in the Dutch media regarding the desirability of offering degrees that are taught in English.

6.1 INTRODUCTION

In Chapter 5 we presented a study in which we used data of Dutch and German students in Nijmegen, the Netherlands, to investigate the relationship between study language and lexical development. In Chapter 6 the same data set is analysed, but the focus is different. This time, we will evaluate an argument that is often used by opponents of the use of English in higher education: that studying in an L2 would be detrimental to content learning. For example, in the Dutch media it is regularly argued that lectures and classroom interactions are not of the same quality when they are held in L2 English as compared to L1 Dutch (e.g., Hermans, 2017; Huygen, 2017; Kleinjan, 2017). This may influence how much students can learn from lectures and classroom interactions. In addition, De Groot (2017, p. 14) argues that the higher mental load that L2 users experience is likely to negatively impact information processing and knowledge transfer.

According to Macaro, Curle, Pun, An and Dearden (2018; see the Introduction to Chapter 5 for a description of this review), the earliest study to investigate such claims actually comes from the Netherlands. In her dissertation, Vinke (1995) investigated Dutch lecturers' and students' perceptions of using Dutch versus English as the language of instruction. She also compared lecturers' teaching behaviour in both languages, as well as how much students learned from listening to a Dutch versus English lecture.

A group of 131 lecturers (out of 245) responded to Vinke's (1995) questionnaire, which targeted the lecturers' perceptions of teaching in Dutch versus English. About 60% of them found the experience of teaching in both languages to be roughly comparable (Vinke, 1995, p. 76). Still, the results on some statements in the questionnaire stood out. For example, the majority of lecturers said they needed (much) more time to prepare courses in English, and felt less capable to express themselves, or to express something in a different way. About half of the respondents also felt less able to improvise in English than in Dutch. Furthermore, about half of the respondents assessed their own teaching quality to be lower in English (Vinke, Snippe & Jochems, 1998).

Vinke (1995) also videotaped 16 Dutch lecturers when they were teaching in Dutch and English. Their teaching behaviours were quite similar, except there was a little more redundancy in the Dutch lectures, and a little more interaction in the English lectures. However, four out of seven teaching behaviours under investigation were rated less favourably by observers[1] when the teaching was in English. These behaviours were body movement, variation in intonation and speed of delivery, verbal fluency and the use of vague terms. No difference was found in the consultation of notes, the use of visual support and the use of gestures. In short, Vinke found evidence that lecturers perceive some aspects of teaching in L2 English to be different (and more difficult) than in L1 Dutch. She also found that the language of instruction affects lecturers' teaching behaviour, as rated by observers.

[1] The observers all had a professional background in education. No information on their language background is provided, but since they were working in the Netherlands it is likely that the observers were Dutch native speakers.

Moving from Dutch lecturers to Dutch students, another question is whether studying in L2 English affects learning outcomes. To investigate this, Vinke (1995) let Dutch students watch a video-taped lecture either in Dutch or in English. The two lectures were given by the same Dutch lecturer, and contained the same content. Afterwards, the students answered true/false questions in Dutch about the content of the lecture they had just seen. Out of 30 questions, the 34 students who had seen the lecture in Dutch on average scored 22.3 points, and the 34 students who had seen the lecture in English on average 21.0 points. This difference amounted to a small-to-medium effect size of $d = 0.45$, and was significant. The students in both conditions had been matched in terms of general academic ability.

Thus, Vinke's (1995) last study showed a detrimental effect of studying in L2 English when it comes to content learning, although the effect was relatively small. This outcome contrasts with that of Vander Beken and Brysbaert (2018), who let native speakers of Dutch (university students) read texts in Dutch and English. After reading, the students answered true/false recognition questions in the same language they had read the text in. No significant effect was found. Potentially, this between-study difference could be explained by the fact that in Vinke the English was spoken by an L2 speaker, whereas the English texts in Vander Beken and Brysbaert were (presumably) written by a native speaker. Another potential explanation is that Vinke tested all students in Dutch, including the ones who had listened to the lecture in English. This discrepancy between the study language and the test language may also have impacted the outcomes. Finally, if the participants in Vander Beken and Brysbaert had struggled to understand the English text, they could have gone back and read it again. This was not possible for the participants in Vinke, who only had one chance to listen to the lecture. For completeness, we note that Vander Beken and Brysbaert not only had their participants answer recognition questions, but also had them write a summary of the texts they had just read. In the summaries, they found better scores in the Dutch condition. The fact that this difference had not been found for recognition suggests that in this particular study it was not L2 comprehension that was difficult for the students, but rather L2 writing.

Macaro et al. (2018) provide a review of studies that target the effects of English-medium instruction (EMI) on content comprehension and learning. In this paragraph, we will summarise the relevant parts of this review. Considering only descriptive statistics (i.e., not using any statistical tests), Hellekjaer (2010) suggested that students' listening comprehension is lower in L2 English lectures than in L1 Norwegian or L1 German. Listening comprehension was measured through a survey with questions about the lecturer's use of English, and about the student's ability to understand the lecturer's line of thought (among other things). Joe and Lee (2013) did not find an effect of study language on listening comprehension, which they measured as medical students' understanding of a medical lecture in a quiz. However, the content of Joe and Lee's L1 Korean lecture differed from that of their L2 English lecture, and Macaro et al. point out that it is unclear whether the lectures had been matched in difficulty. Dafouz, Camacho and Urquia (2014) found no difference in the grades obtained by Spanish students with and without EMI. Macaro et al. state that it is unclear how the tests in this study

were designed, and whether it was controlled how much English was used in the EMI classes (if the use of English was rare, then it is no surprise that the grades did not differ between the two groups). Tatzl and Messnarz (2013) let one group of engineering students read and write a physics exam in L1 German and the other group in L2 English. The scores were not significantly different, but Macaro et al. question the reliability of this finding since there are no data to show that the two groups were equivalent to begin with (e.g., in their knowledge of physics).

In summary, the outcomes summarised in Macaro et al. (2018) are mixed, and coupled with uncertainty as to whether the conditions across the included studies were truly comparable. This led Macaro et al. to conclude that the evidence on the relationship between EMI and content learning is inconclusive. Therefore, Chapter 6 addresses the question of whether study language influences content learning, or more precisely, whether it influences study success in a slightly broader sense, operationalised as obtained ECs, obtained grades, and drop-out rates. This was done by comparing students who followed the exact same programme in either Dutch or English, which should make the groups more comparable than in previous research.

6.1.1 Comparing study success between existing groups

We compared the study success of four groups of first-year students in the Psychology programme of Radboud University in Nijmegen, the Netherlands. This part of the design is identical to that of the study described in Chapter 5, and will be briefly summarised here. We contrasted two nationalities (Dutch and German) and two study languages (Dutch and English). We will use the term *track* for the language in which someone studies. Thus, there were four groups of students: Dutch students in the Dutch track, Dutch students in the English track, German students in the Dutch track, and German students in the English track.

The students were free to choose the language in which they wanted to study, and their nationality was a given. As a result, it is possible that there were pre-existing differences between the four groups which may have influenced their study success. For example, it is conceivable that more intelligent and/or more highly motivated Dutch students are more likely to choose the L2 English track rather than the L1 Dutch track, as the former could be considered more challenging or prestigious. If this is true, it can impact our analysis of study success. For this reason, we first investigated whether there were any pre-existing differences between the Dutch students in the Dutch and English tracks by examining their mean high school exam grade, which is a known predictor of academic achievement in psychology students (e.g., De Koning, Loyens, Rikers, Smeets & Van der Molen, 2012). We did not find such a difference between the two Dutch groups, indicating that they seemed to be comparable in terms of their general academic ability. The German students' high school grades unfortunately were not available, and therefore we could not include them in this analysis of potential pre-existing differences. We return to this issue in the Discussion.

In addition to a relationship between academic achievement and high school grades, a positive relationship between academic achievement and proficiency in the study language has been established (e.g., De Koning et al., 2012; Fonteyne, Duyck & De Fruyt, 2014; Zijlmans, Neijt & Van Hout, 2016). Therefore, we also compared the Dutch students in both tracks on their high school exam grade for English, since even in the Dutch track many of the reading materials were in English. We found that the Dutch students in the English track had scored significantly better on their English high school exam than the Dutch students in the Dutch track. Because such a pre-existing difference between the groups is undesirable, we matched the Dutch students in both tracks on their English high school grade. We did this by removing students from the bigger group (i.e., from Dutch in the Dutch track) until the average high school grade for English in the two groups was the same. This matching procedure is described in detail in the Results section (6.3.2.2). In this way, we ensured as much as possible that any effects of study language on study success would not have been caused by underlying differences in English proficiency between the two groups. It did not seem necessary to also match the students on their Dutch exam grades, since the Dutch language played no role in the English track.

After this correction for pre-existing differences between the two Dutch groups, we investigated whether the four groups differed on several measures of study success. Our use of the term *study success* encompasses both content learning, operationalised through the grades and the number of European Credits (EC) that the students obtained in their courses, as well as whether or not the students dropped out of the Psychology programme. We expected to find superior outcomes for those students who studied in their L1 (i.e., the Dutch students in the Dutch track) as compared to the other three groups. We were also interested in other contrasts, such as the comparison between German students in both tracks, although we had no specific hypotheses for those contrasts.

6.1.2 The relationship between lexical richness and study success

The above-described analysis of the relationship between study language and study success took place at the group level. However, within the groups the students likely also differed in their study language proficiency. We already referred to several studies that showed a positive relationship between students' proficiency in their study language and their academic achievement (see above), but whether, or to what extent, this relationship also holds for lexical richness as an aspect of language proficiency is mostly unknown. Lexical richness concerns how advanced someone's productive vocabulary is (i.e., the words that someone produces in writing or speaking). This includes the diversity of words that someone uses, as well as how difficult or rare these words are. In Chapter 5, we found that the students' lexical sophistication scores (i.e., the proportion of sophisticated words) were significantly related to their grades on open exam questions, but their lexical density scores (i.e., the proportion of content words) and lexical variation scores (i.e., the number of different words) were not. In contrast, Lemmouh (2008) did not find a significant relationship between students' essay

grades and a lexical richness measure called *Lexical Frequency Profile* (LFP), which seems to be similar to lexical sophistication. However, Lemmouh (2008) did find that LFP scores were significantly related to overall course grades.

Douglas (2010) also investigated the relationship between lexical richness and study success, although indirectly. He found that measures of lexical richness predicted students' performance on the *Effective Writing Test* (EWT), which in turn predicted several measures of study success. The EWT is a test employed by the University of Calgary to determine whether aspiring students' academic writing competency is sufficient for admission to the university. EWT scores were positively related to students' cumulative undergraduate grade point average (GPA), and negatively to the number of courses attempted but not passed. When its predictive power was combined with the students' English high school grade, EWT scores were negatively related to the number of semesters students were enrolled at university, and their 'academic standing' (i.e., the number of incidences of academic probation, and how often students were required to withdraw from their study programme). However, the outcomes from Douglas (2010) do not inform us on the strength of the direct relationship between lexical richness and study success.

In short, there are only very few studies on the relationship between lexical richness and study success, and their outcomes point in different directions. Therefore, our last research question concerned this relationship. We investigated whether students who scored higher on the three measures of lexical richness which were also used in Chapter 5 (based on Lu, 2012) and are described above, obtained more ECs, obtained higher grades, and dropped out less often.

6.1.3 Research questions

The following three research questions were addressed in Chapter 6:

1. Are the 'better' Dutch students (i.e., those with higher high school exam grades) more likely to choose the L2 (English) rather than the L1 (Dutch) track?
2. Is there a relation between nationality, study language and study success?
3. Is there a relation between students' lexical richness in their study language, and their study success?

6.2 METHODS

6.2.1 Participants

The data set with 614 participants, as described in Chapter 5, was analysed again. This data set consisted of all the Dutch and German first-year psychology students at Radboud University in the academic year 2016-2017, except for four students who had opted out of participation. Students with a second nationality had been excluded from the data set. To answer Question 1, a subset of this data set was extracted containing the Dutch students for whom VWO high school grades were registered by the university (VWO is the Dutch secondary education type

which prepares students for university). These grades were available for 213 out of the 236 Dutch students. For Questions 2 and 3, we reverted back to the full data set of 614, and then excluded 31 students. They had received an exemption for one or more courses, and/or had also enrolled in extra courses on top of the regular work load. These students were excluded because it is complicated to investigate the relationship between study language and study success when the students in the to-be-compared groups do not follow the same study programme. The demographic information of the remaining 583 students is shown in Table 1.

Table 1. Demographic information of the 583 students in the analysis of study success.

Nationality	Track	*n*	% female	Mean age (*SD*)	Age range
Dutch	Dutch	172	80%	19.52 (2.64)	17 – 47
	English	36	67%	19.17 (1.20)	17 – 23
German	Dutch	50	76%	20.86 (1.88)	18 – 26
	English	325	68%	20.65 (1.92)	17 – 30
	Total	583	72%	20.24 (2.20)	17 – 47

Note. 'Age' refers to the students' age on 26 October 2016, the day of the first of the three exams.

We used various subsets of this data set to answer Questions 2 and 3. Between these subsets, the percentage of female students and student age were comparable to the data in Table 1. As previewed in the Introduction and explained in more detail in the Results section (6.3.2.2), Question 2 was answered with a subset of the data in which the Dutch students in the two groups (i.e., the two tracks) had been matched on their English high school grades. This way, we could examine the effect of Group while controlling for pre-existing group differences in English proficiency and overall high school grades. This was not necessary for Question 3, because the Group variable was no longer of primary interest, and could now be used to take out the variance that was due to pre-existing differences between the groups. We then investigated whether the lexical richness variables could explain the remaining variance in the data.

For both Questions 2 and 3, the students who had dropped out during the first year were excluded in the analyses of grades and obtained number of ECs. This is because the students who dropped out early, on average achieved a lower number of ECs, simply because they did not take as many courses as the other students. Their impending drop-out may also have impacted their grades, and we did not want our investigation of the effect of Group and of lexical richness to be influenced by this.

The data set that was available for Question 3 was smaller than for Question 2, because not all students had completed all three exams from which the lexical richness scores were calculated (for details, see Chapter 5). Out of the data set with 583 students, the lexical richness measures were available for 305 students. None of these students had dropped out

during the first year. In the Analysis and Results sections, for each question we will indicate again which data set we used.

6.2.2 Ethics and data handling

See section 5.2.2 of Chapter 5.

6.2.3 Data

The same data set was used as in Chapter 5. It consisted of:

- Demographic information
- Hand-written answers to three open exam questions, together with the grades that the course lecturer or teaching assistant had assigned to those answers
- The grades and the number of ECs that the students had obtained for the 13 courses that make up the programme of the first study year of Psychology
- The students' high school exam grades; these were available for a subset of the Dutch students

6.2.3.1 Measures of study success

We quantified study success in four different ways. The first was the number of ECs the students obtained (variable name: *Number of ECs*; as in previous chapters, variable names will be written with a capital letter). This variable is directly related to whether the students passed or failed each of the 13 courses that make up the first year of the Psychology programme (variable name: *Passing a course*). This is because all courses are worth a fixed number of ECs, typically between three and six (one EC represents a work load of 28 hours). If a student passes a course, he/she obtains all the ECs which that course is worth. If a student fails, he/she obtains no ECs. A standard one-year study programme covers 60 ECs, therefore the students could at most obtain 60 ECs. The binary variable Passing a course had at most 13 observations with the value of 0 (fail) or 1 (pass). Some students did not enrol in all first-year courses, for example for health reasons or personal circumstances, and for them there would be less than 13 observations. We mention both Number of ECs and Passing a course, which are two sides of the same coin. This is because we explored various statistical models that took dependent variables in different forms, sometimes as one summary statistic (e.g., Number of ECs), and sometimes as repeated measures (e.g., Passing a course). More details are given in the Analysis section and in Appendix A.

Second, we had access to the grades that the students obtained in the 13 courses. In the Dutch system, grades are given on a 1-10 scale, where a 6 or higher is needed to pass. In some courses in our data set, the lecturers only awarded an actual grade when students passed, and a fail otherwise. In those cases, we converted the fail label to a grade of 4, as this is often regarded as the prototypical fail grade. As above, these repeated grade measures (variable name: *Grade*) had an accompanying summary score, in this case the mean (variable name:

Mean grade). In calculating Mean grade, each individual grade was weighted by the number of ECs the course was worth, as courses with more ECs require more work and therefore arguably are more important. If a student did not enrol in a course at all (rather than taking part but failing), this course was not included in the calculation of the mean grade for that particular student.

The Mean grade variable has two potential disadvantages. First, if a student did not enrol in any courses, or did enrol but did not show up for any exams, Mean grade could not be calculated. This was the case for 28 out of the 614 students, and they were excluded from the analysis that focused on Mean grade. Second, there may be a relationship between students' grades and the number of courses they are enrolled in: It should be easier for students to obtain high grades when they take part in fewer courses or exams, as they have less material to study. At the same time, it is conceivable that students struggling with personal issues both obtain lower grades and enrol in fewer courses, and that highly motived students enrol in more courses. However, such potential relationships between grades and number of courses are not reflected by the Grade and Mean grade variables.

Therefore, we constructed an additional variable in which the students' grades were weighted by the number of ECs the students were enrolled in (regardless of whether or not they passed the respective courses). For each student, we multiplied each obtained grade by the number of ECs the corresponding course was worth, and we computed the sum of all those outcomes. For example, if a student obtained a 7.5 in a course of 5 ECs, and a 4 in a course of 3 ECs, the outcome would be 7.5*5 + 4*3 = 49.5. The conceptual difference of this variable (name: *Weighted grade*) to the Mean grade variable is that Weighted grade is not divided by the number of courses a student is taking. Thus, students who were enrolled in fewer courses received a lower score on this variable, and students who did not enrol in, or pass, any courses scored 0. For this reason, Weighted grade should not be used as a repeated-measures variable with (at most) 13 observations; only by summing the scores, those scores also reflect the number of courses that a student took. This means that students scored high on Weighted grade both if they took many courses and if they obtained good grades.

The final variable of study success was whether or not a student dropped out of the Psychology programme (variable name: *Drop-out*). We had access to three potential outcomes: dropping out during the first year, dropping out in between the first and second year, and continuing into the second year. In the analyses we made a binary distinction between whether or not students dropped out, regardless of when they dropped out. It is likely that some additional students also dropped out during the second year or later, but we have no information on that.

6.2.4 Analysis

6.2.4.1 Investigating pre-existing group differences

Question 1 asked whether 'better' Dutch students (i.e., those with better high school exam grades) are more likely to opt for the English track. Since this question was restricted to

the Dutch students (we could not obtain the high school data of the German students), the independent variable was Track (two levels: Dutch, English). We compared the Dutch students in the two tracks on two dependent variables: their mean high school exam grade over all courses, and their high school exam grade for English. Because the high school grades mostly seemed to be non-normally distributed, we used the non-parametric Wilcoxon rank-sum test with continuity correction to compare means. We also present bootstrapped descriptive statistics so that there is no need to rely on distributional assumptions (see section 6.2.4.3).

6.2.4.2 Predicting ECs and grades from Group

Question 2 asked whether the four groups differed on the four measures of study success (Number of ECs, Mean grade, Weighted grade and Drop-out). The four groups were Dutch students in the Dutch track, Dutch students in the English track, German students in the Dutch track, and German students in the English track. We chose to directly compare study success between the four groups, as opposed to the option of having a two-by-two design with Nationality (Dutch versus German) and Track (Dutch versus English) as the independent variables. We preferred this design because the sample sizes between the four groups were quite different (see Table 1). This means that any potential main effects of Nationality and Track could not be meaningfully interpreted, as was explained in Chapter 5 (section 5.2.4.1).

We explored the use of linear mixed-effects models and analysis of variance (ANOVA) to compare the four groups on the first three dependent variables (the binary variable Drop-out is discussed in section 6.2.4.5). This is detailed in Appendix A and summarised here. All statistics were performed in R (R Core Team, 2018). The model diagnostics of the linear mixed-effects models showed that this technique was not suitable for modelling the variable Passing a course (i.e., the repeated-measures equivalent of Number of ECs). The residual plot showed a positive relationship between the predicted values and the residuals, while no such patterns should be visible in a good model. In addition, more than 5% of the model's predictions fell outside of the theoretical 95% error bounds. Using ANOVA to model Number of ECs was not an acceptable option either, because Number of ECs was very much non-normally distributed in all four groups (the most common score was the maximum score of 60 ECs).

A solution for modelling data sets that cannot be approximated well by parametric models is to use bootstrapping. The basic idea behind this approach is that rather than making assumptions about the shape of the distribution of the data that one is modelling, this distribution is actually estimated from the data. Field and Wilcox (2017) stress that the use of such models is much preferred over the use of linear models whose assumptions are not met. Therefore, we performed an ANOVA with bootstrapping on Number of ECs. The exact procedure for conducting an ANOVA with bootstrapping is explained in the next section.

It was also necessary to use ANOVA with bootstrapping for Weighted grade, since this variable was non-normally distributed within the groups too. For Mean grade, we had multiple options. The repeated-measures variable Grade could be modelled satisfactorily with a linear mixed-effects model, and the summary score Mean grade could be modelled

satisfactorily with a 'regular', parametric ANOVA. However, in order to preserve consistency between the analyses, we decided to use bootstrapping for modelling Mean grade through ANOVA as well.

6.2.4.3 Robust statistics

The use of bootstrapping is one of the ways in which 'robust' statistics can be obtained, which are statistics that are not dependent on the data being normally distributed. The normality assumption is also relevant for calculating descriptive statistics. For example, for non-normally distributed data, it is recommended to report the median rather than the mean. The reason is that the mean is much more sensitive to shifts from normality (in other words, it is less robust), and might not represent the most typical case (Wilcox, 2005). In this chapter, we will report both means and median values, to provide maximum insight in the data set. Like the mean, the standard deviation (SD) is also a non-robust measure (Högel, Schmid & Gaus, 1994; Wilcox, 2005). To obtain a robust estimate of the variability of the observations, we used bootstrapping. In this procedure, the distribution of a certain measure (e.g., the mean or median) is estimated with data from the study's sample, rather than assumed to be normal (Efron & Tibshirani, 1993). This is done by taking a large number of samples of size *N* (with replacement) from the original study's sample, where *N* equals the original study's sample size.

As an example, we will consider the grand mean and standard deviation of the dependent variable Number of ECs. Say we have a total of 500 observations of this measure. One bootstrap sample thus would consist of 500 observations randomly drawn with replacement from these 500 original observations. Drawing with replacement means that after a value has been drawn it goes back into the virtual 'jar' and can be drawn again. Therefore, some values may end up in the bootstrap sample multiple times, while others may be absent. Then, the mean of the 500 randomly drawn observations is calculated. In our case, we always repeated these procedures 10,000 times. The resulting set of 10,000 means follows a certain distribution. We can use this distribution to calculate robust standard errors (SE) and confidence intervals (CI) of the mean. More specifically, for our calculations we used bias-corrected and accelerated (BCa) CIs (see DiCiccio & Efron, 1996)[2]; this is recommended over simply extracting the middle 95% (if α equals .05) of all values (Wright, London & Field, 2011). SDs were calculated from the SEs.

Bootstrapping can also be used to obtain inferential statistics. Wilcox (2005) developed a procedure for conducting a robust ANOVA with bootstrapping. In general, in ANOVA an *F*-statistic is calculated that represents the ratio between the variance that is explained by the model, and the unexplained variance. This *F*-statistic is compared against the *F*-distribution, which is the probability distribution of the *F*-statistic when the null hypothesis

[2] *Bias-corrected* refers to the correction that is applied when the bootstrap mean is biased, *accelerated* refers to a technique which lets the limits of the CIs converge more quickly (Wright, London & Field, 2011, p. 259).

is true. In robust ANOVA, the *F*-distribution is not assumed to be known (as it is in parametric ANOVA), but rather a new *F*-distribution is obtained through bootstrapping. Generating the bootstrap samples is done in the same way as described above, using a certain number of iterations (we again used 10,000). Critically, in each iteration, the resulting bootstrapped data within each group are then centred around 0 by subtracting the group mean from each data point. Thus, all group means become 0, and therefore do not differ from one another. An *F*-statistic is computed from these data (for the details of this calculation, see Wilcox, 2005, p. 267). The resulting distribution of (say) 10,000 *F*-values is the distribution we can expect when the null hypothesis is true. In a final step, the *F*-statistic from the observed (non-bootstrapped) data is calculated and compared against the *F*-values in the bootstrapped distribution. The *p*-value is calculated as the proportion of values from the bootstrapped distribution that are equal to or bigger than the *F*-value from the actual data.

6.2.4.4 Predicting ECs and grades from lexical richness

Question 3 asked whether students' lexical richness in their study language predicts their study success. The four measures of study success were again the dependent variables (of which Drop-out is discussed in the next section). Group plus the three lexical richness measures were the independent variables; Group was no longer of direct interest in Question 3, but could be used to account for potential variation due to pre-existing group differences. Only after including Group in our model, we looked at the added value of including the three lexical richness measures.

The lexical richness measures were lexical density (LD), lexical sophistication (LS), and lexical variation (LV). LD concerns the ratio of content words (nouns, verbs, adjectives and adverbs) to the total number of words in a text. LS concerns the ratio of sophisticated word types to the total number of word types in a text (each unique word in a text is a word type). Words were considered sophisticated if they did not appear in a list of the 2,000 most frequent words in the language (i.e., in Dutch or English). LV concerns the diversity of the words in the text, and was calculated as the average number of different words in 10,000 randomly drawn 20-word samples. More details can be found in section 5.2.3.3 of Chapter 5. For the present study, we arrived at one value of LD, LS and LV per student by averaging the lexical richness scores over the three exam answers that the students had written.

As with Question 2, the model diagnostics of parametric models we explored for Passing a course (i.e., the repeated-measures equivalent of Number of ECs) did not look good: The residual plot showed a positive trend and most of the predicted data points fell outside of the theoretical 95% error bounds. Details are provided in Appendix B. In this case, modelling Number of ECs with a robust regression with bootstrapping did not seem to provide a solution either because the relationship between the dependent and independent variables may not have been linear, since the number of ECs had an upper bound of 60.

In Appendix B we explain that there were also distributional problems with Weighted grade. By this stage the analysis of Question 2 had already shown us that Weighted grade

did not bring much added value to Mean grade. We therefore decided to limit Question 3 to the two variables with unproblematic distributions: Grade (i.e., the repeated-measures equivalent of Mean grade) and Drop-out (see section 6.2.4.5). As shown in Appendix B, the model assumptions for Grade could be met.

The relationship between Grade and the lexical richness measures was analysed through a series of nested linear mixed-effects models:

1. Grade ~ 1 + Group + (1 | Subject) + (1 | Course)
2. Grade ~ 1 + Group + LD + (1 | Subject) + (1 | Course)
3. Grade ~ 1 + Group + LD + LS + (1 | Subject) + (1 | Course)
4. Grade ~ 1 + Group + LD + LS + LV + (1 | Subject) + (1 | Course)

The syntax of these models should be read as follows. The dependent variable Grade is modelled from an intercept (represented by '1'), a number of fixed effects (written without parentheses), and two random effects, namely random intercepts at the subject and the course level. Each model was compared to the model above it by means of a likelihood ratio test, in order to assess whether the addition of a new variable significantly increased the model's goodness of fit to the data.

6.2.4.5 Investigating drop-out rates

To investigate the effect of Group on Drop-out (Question 2), we ran a logistic regression with Group as predictor. After running this model, we checked the model diagnostics, following Field, Miles and Field (2012, p. 341). They are reported in Appendix A, which shows there was no reason for concern.

To examine the effect of lexical richness on Drop-out (Question 3), we first explored the option of running a hierarchical logistic regression with Group, LD, LS and LV as predictors. This means we started out with a model containing only Group as a predictor, and then created three further models in which LD, LS and LV were added one by one. The increase in model fit between the models was compared with likelihood ratio tests. In this way, we could preserve consistency to the analysis of the effect of lexical richness on Grade. However, the model diagnostics showed that in all of the models, some of the data points exerted too much influence on the regression line. However, we could not simply remove these cases, because between the four models, the problematic cases were not the same.

Therefore, we switched to an alternative way of modelling the data, namely with a forced-entry logistic regression. This means we added all the predictors (Group, LD, LS and LV) at once, and their order was of no concern. We checked the model diagnostics for this model, and identified five outliers that also exerted a disproportionate influence on the model (see Appendix B for the details). We ran the model again with these five cases excluded and compared its outcomes to the outcomes from the original model in which these cases were still included, in order to assess the robustness of our model.

6.2.4.6 Controlling Type-I error rates

As in Chapter 5, we started out with α = .05 and corrected for multiple testing within each research question. Again, we used Nyholt's (2004) software which calculates the significance threshold for a set of dependent variables where the correlation between these variables has been taken into account. This method ensures that Type-I error rates stay at 5%, while it is less conservative than a Bonferroni correction. For example, since Question 1 was evaluated on two dependent variables, a Bonferroni-corrected α would be .0250. However, Nyholt's (2004) significance threshold was .0283 and Li and Ji's (2005) threshold was .0253. Nyholt (2015) states that Li and Ji's threshold is "reportedly more accurate", although it should only be used if it is below Nyholt's (2004) threshold. Thus, in this case we set α to .0253. In the Results section, we present the new α-level along with each research question.

In addition to correcting α for the use of multiple dependent variables, we also corrected α whenever we made multiple comparisons as a follow-up to the robust ANOVAs. This was done with the Benjamini-Hochberg (1995) procedure, in which the *p*-values from all the contrasts are first ordered from smallest to largest. The rank of a *p*-value is *j*, where the smallest *p*-value has rank 1 and the largest has rank *k*. The corrected level of α, per *p*-value, is $j/k * \alpha$.

6.2.4.7 Confidence intervals for percentages

Drop-out rates were reported as a percentage. Confidence intervals around these percentages were obtained with the following formula:

$$p \pm z * \sqrt{\frac{p * (1 - p)}{n}}$$

Here, *p* is the proportion of students dropping out, *n* is the sample size, and *z* is the critical value of *z* for the desired level of confidence. We worked with *z* = 2.33, which corresponds to a confidence level of 98%. This is because α for Question 2 was .0202, which is rounded off to 98% in *z*-tables.

6.3 RESULTS

6.3.1 Question 1: Do the better Dutch students opt for the L2 track?

6.3.1.1 Alpha

Question 1 was evaluated on two dependent variables. Nyholt's (2004) significance threshold was .0283 and Li and Ji's (2005) threshold was .0253. Thus, we set α to .0253.

6.3.1.2 Data set

We used the data of 213 Dutch students whose high school grades were registered in the university system. Students with atypical study programmes were not excluded from the data set for this research question, because we were interested in the students' academic ability and English proficiency *before* they had started their university programmes.

6.3.1.3 Are there pre-existing group differences?

Table 2 shows the average high school exam grades of the Dutch students who later opted for the Dutch versus the English track.

Table 2. High school exam grades of the Dutch students.

High school exam	Track	*n*	Mean (*SD*)	97.47% CI	Median	Range
English	Dutch	174	6.87 (0.84)	6.72 – 7.01	7	5 – 9
	English	39	7.49 (0.75)	7.21 – 7.74	7	6 – 9
Overall (all courses)	Dutch	174	6.69 (0.49)	6.61 – 6.78	6.58	5.89 – 8.50
	English	39	6.78 (0.40)	6.65 – 6.94	6.73	6.08 – 7.85

Note. SDs and CIs were obtained through bootstrapping with 10,000 iterations.

The Dutch students who had opted for the English track had obtained a significantly higher English grade in high school ($W = 4692$, $p < .001$). Hedges' g was 0.75, indicating a medium-sized effect. There was no significant difference in overall high school grades between the students in both tracks ($W = 4024.5$, $p = .07$), although the trend again was in favour of the Dutch students in the English track, with a (very) small effect size of $g = 0.19$.

6.3.2 Question 2: Do study language and nationality affect study success?

6.3.2.1 Alpha

Question 2 was evaluated on four dependent variables. This time, Nyholt's (2004) threshold was .0202 and Li and Ji's (2005) threshold was .0253. Since Li and Ji's estimate was not smaller than Nyholt's, we set α to .0202 for all analyses belonging to Question 2.

6.3.2.2 Data set

In answering Question 1, we had found that the two Dutch groups were unequal from the beginning in terms of their pre-existing knowledge of English, which would be problematic for the rest of the analyses. Contrary to what is often thought, such pre-existing differences cannot simply be controlled for by using analysis of covariance (ANCOVA), because independence of the covariate (here: high school English grade) and treatment (here: Group) effect is a prerequisite of ANCOVA (Field et al., 2012, pp. 464-466). As Field et al. (2012) explain, pre-existing group differences should be resolved in one of two ways: either randomising

participants to experimental groups (which was not an option for us), or matching the groups on the covariate.

Therefore, we conducted a procedure to match the two Dutch groups on their English grade. We used the data set with 583 students (i.e., the data set from which students with atypical study programmes had already been excluded, see the Participants section, 6.2.1). Of these students, 208 were Dutch. In order to match the Dutch students on their English grade, we first excluded all Dutch students for whom high school grades were not available (n = 19). This left 156 Dutch students in the Dutch track and 33 Dutch students in the English track. We then, one by one, removed students from the largest group (i.e., from Dutch in Dutch track) until the average English grade of the two groups was the same to two decimal places. To achieve this, the Dutch students in the Dutch track were first sorted by their English exam grade, from low to high. We then removed the student with the lowest English grade and checked whether the mean English grade now was the same between the two groups. When this was not the case, we removed the student with the next-lowest grade and checked again. After repeating this procedure 62 times, the average English grade in both groups was the same. Thus, a total of 62 Dutch students in the Dutch track were removed from the data set. This left 94 Dutch students in the Dutch track for analysis (and 33 Dutch students in the English track). In the end, the data set now contained 127 Dutch students and 375 German students, a total of 502 students.

Table 3 shows the high school exam grades of the Dutch students in both groups after the matching procedure had been completed. Because α was .0253, we calculated 94.47% CIs. The Dutch students in the two tracks still did not differ significantly on their overall exam grade (W = 1429, p = .50), which the matching procedure had not targeted.

Table 3. High school exam grades of the Dutch students after the matching procedure.

High school exam	Track	*n*	Mean (*SD*)	97.47% CI	Median	Range
English	Dutch	94	7.45 (0.55)	7.32 – 7.57	7	7 – 9
	English	33	7.45 (0.74)	7.15 – 7.73	7	6 – 9
Overall (all courses)	Dutch	94	6.86 (0.48)	6.75 – 6.97	6.77	6.10 – 8.15
	English	33	6.77 (0.39)	6.64 – 6.95	6.67	6.18 – 7.85

Note. SDs and CIs were obtained through bootstrapping with 10,000 iterations.

For the analysis of ECs, Grade and Weighted grade, we removed 71 students who had dropped out during the first year from this matched data set. As explained earlier, students who drop out most likely obtain fewer ECs, and lower grades. This could obscure the analysis of the effect of Group. In the resulting data set of 431 students, the two Dutch groups still did not differ significantly on English grade (W = 1381.5, p = .82) or mean high school grade (W = 1182, p = .31). For the analysis of Drop-out, of course we did not exclude those students who dropped out from the matched data set.

6.3.2.3 Number of ECs

Table 4 contains the descriptive statistics for Number of ECs that was obtained in the four groups. The outcome of the robust ANOVA was non-significant ($F(3, 427) = 3.24$, $p = .06$) against an α-level of .0202.

Table 4. Descriptive statistics for Number of ECs obtained.

Group	*n*	Mean (*SD*)	97.98% CI	Median	97.98% CI	Range
Dutch in Dutch track	87	54.26 (9.85)	51.00 – 56.23	60	54 – 60	11 – 60
Dutch in English track	31	49.52 (15.78)	40.87 – 54.61	54	42 – 54	0 – 60
German in Dutch track	40	49.25 (13.73)	43.05 – 53.38	54	46.06 – 55.50	7 – 60
German in English track	273	50.29 (15.28)	47.93 – 52.22	60	54 – 60	0 – 60
Total	431	50.94 (14.42)	49.14 – 52.44	60	54 – 60	0 – 60

Note. SDs and CIs were obtained through bootstrapping with 10,000 iterations.

6.3.2.4 Mean grade

Descriptive statistics for the Mean grade variable are given in Table 5.

Table 5. Descriptive statistics for Mean grade.

Group	*n*	Mean (*SD*)	97.98% CI	Median	97.98% CI	Range
Dutch in Dutch track	87	7.23 (0.82)	7.02 – 7.44	7.25	6.99 – 7.46	5.18 – 9.02
Dutch in English track	31	6.68 (1.08)	6.14 – 7.07	6.84	6.44 – 7.09	3.77 – 8.60
German in Dutch track	40	6.73 (0.97)	6.33 – 7.06	6.85	6.34 – 7.19	3.88 – 8.79
German in English track	270	6.91 (1.06)	6.76 – 7.05	7.10	6.85 – 7.22	3.41 – 8.81
Total	428	6.94 (1.01)	6.82 – 7.05	7.10	6.96 – 7.17	3.41 – 9.02

Note. SDs and CIs were obtained through bootstrapping with 10,000 iterations. The total *N* is slightly lower as compared to the *N* in Tables 4 and 7, because a few students did not take any exams, and Mean grade could not be computed for them.

This time, the outcome of the robust ANOVA was significant ($F(3, 424) = 4.62$, $p = .007$). The outcomes of robust post-hoc tests are shown in Table 6. After correcting α for multiple comparisons, these tests showed that the Dutch students in the Dutch track scored significantly better than all other groups. In contrast, the other three groups did not significantly differ from one another.

Table 6. Pairwise comparisons between all groups on Mean grade.

Contrast	Mean difference *(SE)*	97.98% CI	*p*	Corrected α
Dutch in Dutch track – Dutch in English track	0.55 (0.22)	0.09 – 1.06	**.005**	3/6 * .0202 = .010
Dutch in Dutch track – German in Dutch track	0.50 (0.18)	0.08 – 0.91	**.004**	2/6 * .0202 = .007
Dutch in Dutch track – German in English track	0.32 (0.11)	0.07 – 0.57	**.0028**	1/6 * .0202 = .0034
Dutch in English track – German in Dutch track	-0.05 (0.25)	-0.64 – 0.50	.85	6/6 * .0202 = .020
Dutch in English track – German in English track	-0.23 (0.21)	-0.74 – 0.23	.27	4/6 * .0202 = .013
German in Dutch track – German in English track	-0.18 (0.17)	-0.57 – 0.21	.29	5/6 * .0202 = .017

Note. Alpha levels were corrected with the Benjamini-Hochberg (1995) procedure. Significant *p*-values are printed in bold.

6.3.2.5 Weighted grade

Table 7 contains the descriptive statistics for the Weighted grade variable (i.e., the sum of each grade times its corresponding number of ECs, see section 6.2.3.1).

Table 7. Descriptive statistics for Weighted grade.

Group	*n*	Mean (SD)	97.98% CI	Median	97.98% CI	Range
Dutch in Dutch track	87	428 (63)	410 – 442	435	412 – 447	109 – 541
Dutch in English track	31	393 (87)	342 – 419	408	380 – 426	57 – 516
German in Dutch track	40	391 (79)	357 – 416	395	377 – 430	113 – 526
German in English track	273	401 (93)	386 – 413	426	411 – 433	0 – 527
Total	431	405 (88)	394 – 414	426	414 – 431	0 – 541

Note. SDs and CIs were obtained through bootstrapping with 10,000 iterations.

The outcome of the robust ANOVA again was again significant against the α of .0202 ($F(3, 427) = 4.16$, $p = .019$). We provide pairwise comparisons in Table 8. The Dutch students in the Dutch track significantly outperformed both groups of German students, but not the Dutch students in the English track. Again, there were no significant differences between the other three groups.

Table 8. Pairwise comparisons between all groups on Weighted grade.

Contrast	Mean difference *(SE)*	97.98% CI	*p*	Corrected α
Dutch in Dutch track – Dutch in English track	36 (17)	0.60 – 78	.018	3/6 * .0202 = .010
Dutch in Dutch track – German in Dutch track	37 (14)	5 – 72	**.0064**	2/6 * .0202 = .0067
Dutch in Dutch track – German in English track	27 (9)	6 – 47	**.002**	1/6 * .0202 = .0034
Dutch in English track – German in Dutch track	1 (20)	-48 – 45	.94	6/6 * .0202 = .020
Dutch in English track – German in English track	-9 (17)	-51 – 26	.62	5/6 * .0202 = .017
German in Dutch track – German in English track	-10 (14)	-44 – 21	.48	4/6 * .0202 = .013

Note. Alpha levels were corrected with the Benjamini-Hochberg (1995) procedure. Significant *p*-values are printed in bold.

6.3.2.6 Drop-out

Table 9 shows the drop-out rates per group. We compared the total drop-out rates between the groups. In other words, we made no distinction between students who had dropped out during and after the first year, because the sample sizes would have gotten very small.

Table 9. Percentage of students who dropped out in each group.

Group	*n*	During year 1	After year 1	Total	98% CI Total
Dutch in Dutch track	94	7.45%	11.70%	19.15%	9.69 – 28.61
Dutch in English track	33	6.06%	15.15%	21.21%	4.63 – 37.79
German in Dutch track	50	20.00%	14.00%	34.00%	18.39 – 49.61
German in English track	325	16.00%	17.23%	33.23%	27.14 – 39.32
Total	502	14.14%	15.74%	29.88%	25.12 – 34.64

Table 9 suggests that the German students dropped out more often than Dutch students. To investigate this observation based on visual inspection, we ran a logistic regression with Group as predictor. Table 10 shows all six contrasts, including the corrected level of α. None of the contrasts were significant at this α-level. Thus, we found no significant difference in drop-out rates between the four groups.

Table 10. Planned contrasts between all groups.

Contrast	*b*	*SE*	*z*	*p*	Corrected α
Dutch in Dutch track – Dutch in English track	0.13	0.50	0.26	.80	5/6 * .0202 = .017
Dutch in Dutch track – German in Dutch track	0.78	0.40	1.96	.05	2/6 * .0202 = .007
Dutch in Dutch track – German in English track	0.74	0.29	2.58	.01	1/6 * .0202 = .003
Dutch in English track – German in Dutch track	0.65	0.52	1.25	.21	4/6 * .0202 = .013
Dutch in English track – German in English track	0.61	0.44	1.39	.16	3/6 * .0202 = .010
German in Dutch track – German in English track	-0.03	0.32	0.11	.91	6/6 * .0202 = .020

Note. Alpha levels were corrected with the Benjamini-Hochberg (1995) procedure.

6.3.2.7 Summary

In Question 2 we examined whether the four groups significantly differed on the dependent variables of Number of ECs, Mean grade, Weighted grade, and Drop-out. No significant effects were detected for Number of ECs and Drop-out. On Mean grade and Weighted grade, the Dutch students in the Dutch track significantly outperformed the German students in the Dutch and in the English track. Only on Mean grade, the Dutch students in the Dutch track also scored significantly higher than the Dutch students in the English track.

6.3.3 Question 3: Does lexical richness affect study success?

6.3.3.1 Alpha

Question 3 was evaluated on two dependent variables. For these variables, Nyholt's (2004) threshold was 0.0273 and Li and Ji's (2005) threshold was .0253. Thus, for Question 3 α was set to .0253.

6.3.3.2 Data set

For Question 3, we used the full, unmatched data set with 583 students again. The reason is that in this case we could use the Group variable to explain the variance that was caused by the pre-existing group differences (we were no longer trying to estimate the effect of Group, because that had already been done in Question 2). Only then, we checked whether the inclusion of the three lexical richness variables had added value to the model. This seemed the better option as compared to using the matched data set, because we could include 81 (62+19) more cases. From this full data set, we extracted the data of 305 students who had completed all three exams (which was necessary for calculating the lexical richness measures). None of these students had dropped out during the first year. Thirty-seven students dropped out in the summer between the first and second year, and 268 students continued into the second year.

6.3.3.3 Grade

Table 11 shows the outcomes of the model comparisons, where each model was compared to the model directly above (only the 'BIC' column concerns the model itself, and not the comparison to the above model). As can be seen, none of the lexical richness variables contributed significantly to the model's fit to the data. Furthermore, the Bayesian Information Criterion (BIC) scores of the models in which the lexical richness variables were included were even higher than that of the model in which only Group was included as a fixed effect. This also indicates that these models are less good (because good models have low BIC scores).

Table 11. Model comparisons to examine the relationship between lexical richness and Grade.

Model	BIC	χ^2	*df*	*p*
1 + Group + (1 | Subject) + (1 | Course)	10773			
1 + Group + LD + (1 | Subject) + (1 | Course)	10778	3.32	1	.07
1 + Group + LD + LS + (1 | Subject) + (1 | Course)	10786	0.08	1	.77
1 + Group + LD + LS + LV + (1 | Subject) + (1 | Course)	10794	0.02	1	.89

Note. LD = Lexical density, LS = Lexical sophistication, LV = Lexical variation.

6.3.3.4 Drop-out

As described in the Methods section (6.2.4.5), we ran the Drop-out model twice, once on the full data set and once with five influential cases excluded. The outcomes of the two models were fairly different, as can be seen from Tables 12 and 13. At our established α-level of .0253, the effect of LD was significant and much stronger after excluding five outlying and influential cases, but not when including these cases. The *b*-estimates of LS and LV in the model with five cases excluded were about half the size of what they were when calculated on the full data set, but in this case the (non-)significance was not affected.

Table 12. Predicting Drop-out from Group and the lexical richness variables on the full data set.

	b	*SE*	*z*	*p*
Intercept (Dutch in Dutch track)	-2.17	0.37	-5.88	**< .001**
Dutch in English track	0.42	0.76	0.55	.58
German in Dutch track	0.77	0.65	1.19	.24
German in English track	0.07	0.52	0.13	.90
LD	-12.71	5.93	-2.14	.03
LS	3.68	4.88	0.76	.45
LV	-0.18	0.34	-0.54	.59

Note. LD = Lexical density, LS = Lexical sophistication, LV = Lexical variation. Significant *p*-values are printed in bold.

Table 13. Predicting Drop-out from Group and lexical richness variables, having excluded five outliers and influential cases.

	b	*SE*	*z*	*p*
Intercept (Dutch in Dutch track)	-2.17	0.39	-5.54	**< .001**
Dutch in English track	-0.14	0.90	-0.16	.87
German in Dutch track	0.12	0.84	0.14	.89
German in English track	-0.29	0.57	-0.51	.61
LD	-20.41	6.62	-3.08	**.002**
LS	1.52	5.50	0.28	.78
LV	-0.10	0.38	-0.27	.79

Note. LD = Lexical density, LS = Lexical sophistication, LV = Lexical variation. Significant *p*-values are printed in bold.

6.3.3.5 Summary

None of the lexical richness measures was a significant predictor of Grade. LD significantly predicted Drop-out (better LD scores were associated with lower drop-out rates), but only after five outlying and influential cases had been excluded from the data set.

6.4 DISCUSSION

6.4.1 Do the better Dutch students opt for the L2 track?

There was no significant difference in overall high school exam grades between the Dutch students who chose to study psychology in Dutch and in English. Thus, it seems that the general academic abilities (as far as they are reflected in high school grades) of the two groups were the same. On the other hand, the Dutch students who chose to study in English had obtained a significantly higher English grade on their high school exam. This is not surprising, as Dutch students with relatively weak English skills would not prefer to study in a completely English programme. Still, even in the Dutch track many of the study materials (e.g., the text books) were in English. Therefore, the fact that we found a significant pre-existing difference in English skills between the two groups was relevant when comparing study success between those groups.

6.4.2 Do study language and nationality affect study success?

In our second research question, we investigated whether there were differences in study success between Dutch students in the Dutch track (i.e., students who studied in their L1), Dutch students in the English track (i.e., students who studied in an L2), and German students in the Dutch and English track (i.e., students who studied in an L2, and outside their home country). To be able to compare study success between the two groups of Dutch students without their English skills being a confounding variable, we matched the students in the two groups on their English exam grade.

We did not have access to the high school grades of the German students, which means we cannot be equally sure about the comparability of the German students in the two tracks. On the one hand it is conceivable that the German students in the Dutch track could have been more motivated and/or have had a higher language aptitude, because German students generally would have had much less experience with Dutch as compared to English before coming to the Netherlands. On the other hand, our findings in Chapter 5 also suggested the possibility that the German students in the Dutch track may have been (near-)native speakers of Dutch. Either way, in the case of a significant difference between the two German groups, we could not be completely sure what is the underlying cause of these differences. Thus, in future studies it would be desirable to have access to German students' grades and language background information. For the current study, the comparisons involving the German groups nevertheless are still interesting for university lecturers and policy makers.

6.4.2.1 Number of ECs

We did not detect a significant effect of Group on the Number of ECs, although descriptively the Dutch students in the Dutch track outperformed the three other groups (and the *p*-value was .06, at an α-level of .0202). Please recall that students earn ECs when obtaining a grade of 6 or higher. The mean grade obtained in all four groups was amply above this cut-off point, the lowest mean grade being 6.68 for Dutch students in the English track. Thus, Number of ECs may have been too coarse a measure to distinguish between the students in the four tracks.

On the other hand, Number of ECs does not only depend on the mean grade, but also on the number of courses a student was enrolled in. This was our motivation for constructing the Weighted grade variable, whose outcomes are discussed below. In any case, Number of ECs continues to be very relevant because it underlies policies such as the *Bindend Studieadvies* (literally in English: *Binding Study Advice*). This policy states that students are only allowed to continue to the second year of study when they have obtained a certain number of ECs in the first year. The details vary per Dutch university and per programme, but psychology students at Radboud University currently have to obtain 42 ECs to be admitted to the second year of study.

From the societal point of view, the present outcome is not bad. It shows that there is no immediate reason for concern with regard to any of the four groups under investigation. It is still conceivable that other groups, such as non-Dutch and non-German international students, might underperform relative to the current participants. Since the number of international students in the Netherlands increases each year (from 41,201 in 2006 to 81,392 in 2016; Huberts, no date), we recommend that other groups are included in future enquiries of study success as well.

6.4.2.2 Mean grade

The Dutch students in the Dutch track obtained the highest mean grades, significantly outperforming all other groups. At least for the comparison between the Dutch students in the Dutch and in the English track, differences in high school grades cannot be responsible for this result, because we had created subgroups that were matched on high school grades. This outcome shows that there indeed seems to be an advantage of studying in the L1, even when a lot of the study materials are in English. This raises the question of what exactly caused the L1 advantage. It may be that the students' language skills enabled them to better study, remember and reproduce the material, but it is also possible that the advantage came from listening to a lecturer who was teaching in his native language. Our finding is in line with Vinke (1995), who found that Dutch students scored better on true/false content questions after watching a lecture in Dutch rather than English (the lecturer was a native speaker of Dutch). However, that study cannot disentangle the two potential explanations either.

Some of the other studies that we mentioned in the Introduction do provide evidence for one of the potential explanations. At the student level, Hellekjaer (2010) and Dafouz et al. (2014) found evidence that students' listening comprehension is lower in the L2. At the lecturer level, a study in Vinke's (1995) dissertation showed that native Dutch lecturers believe their teaching is of lower quality in English as compared to Dutch. Dutch lecturers were also rated less favourably by observers when they were teaching in English rather than Dutch.

Of course, the two explanations are not mutually exclusive. In the context of Dutch higher education, it would be interesting to determine their relative importance. This would yield more insight in the question of whether students' academic achievements benefit more from offering English language training to students and/or to lecturers. An interesting experiment could be to let Dutch students attend a lecture that is either given by an L1 Dutch speaker in Dutch or an L1 English speaker in English, and have them answer content questions afterwards. This would ensure that any effects would be due to the students' language proficiency, and not the nativeness of the lecturer. The comparability of the materials in the Dutch and English condition should be strictly controlled.

If we were to invest in English language training for students, the next question is which aspects of English such training should focus on. There is a wide range of research showing that language proficiency is positively related to academic achievement (e.g., De Koning et al., 2012; De Wachter, Heeren, Marx & Huyghe, 2013; Fakeye & Ogunsiji, 2009; Fonteyne et al., 2014; Van der Westen & Wijsbroek, 2011; Zijlmans et al., 2016). However, none of these studies compared the relative importance of different aspects of language proficiency. For example, should one invest more in training vocabulary knowledge, grammar, general reading skills, or something else? Nevertheless, there are other studies that have focused on L2 reading comprehension, which at least seems to be a central aspect of studying in an L2 (although it should be noted that Fonteyne et al. (2014) did not find that L1 reading comprehension predicted academic success). Jeon and Yamashita (2014) summarised this

research in a meta-analysis, which showed that, among other things, L2 grammar knowledge, L2 vocabulary knowledge and L2 decoding (i.e., converting letters to sounds) were strongly correlated with L2 reading comprehension. Thus, all of these domains may be an interesting target of language training for students. Future research should examine whether these and other correlates of L2 reading comprehension can also predict academic achievement. If we were to invest in English language training for lecturers, the same question applies. Which aspects of language proficiency (e.g., vocabulary, pronunciation, fluency) are most strongly related to teaching skills?

A completely different conclusion that one could draw from the same data is that the use of English in Dutch higher education should be restricted wherever possible. Irrespective of the outcomes of the current study, many people already argue for such a change, and for various reasons (e.g., Huygen, 2017; Teuling, 2017; Vasterman, 2017). Of course, such a decision would depend on many other factors as well, such as the wish to recruit non-Dutch speaking students or lecturers, or the desire to create an international classroom.

However, before implementing any new policies on the basis of the outcomes of this study, we should consider how much importance should be attached to the magnitude of the effects that we found, regardless of their significance. The largest difference in Mean grade between the four groups was at 0.55 points on a 1-10 scale (this difference was found between the Dutch students in the Dutch track and the Dutch students in the English track), corresponding to a medium effect size of $g = 0.62$. Such a grade point difference may or may not warrant the costs of offering English training to students and/or lecturers, and may or may not warrant radical changes in language policy in Dutch higher education. In addition, longitudinal research is needed to investigate how the differences in grade between the four groups develop over the study period of three years (i.e., the duration of a bachelor's degree).

6.4.2.3 Weighted grade

A significant effect of Group was also found on the Weighted grade variable, although in this case the Dutch students in the Dutch track only significantly outperformed the two German groups. However, the trend that was visible was in the expected direction, with the Dutch students in the Dutch track also outperforming the Dutch students in the English track with $g = 0.50$, an effect of medium strength. The small sample size of Dutch students in the English track, coupled with our strict correction of α, may have prevented this effect from reaching significance (the *p*-value was .018, with an α-level of .010). Potential explanations for the native advantage would be the same as those discussed above for Grade. It is good to know that the Weighted grade variable, which we had constructed in order to resolve potential issues associated with the Mean grade variable, in fact turned out to be less sensitive. In future studies it therefore seems unnecessary to include this measure again.

6.4.2.4 Drop-out

The descriptive statistics showed quite substantial differences between the drop-out rates of Dutch and German students, but these differences did not reach significance. We should point out that our α-level of .003 was very strict, and as a result a *p*-value as low as .01 did not reach significance. It therefore seems worthwhile to continue monitoring drop-out rates between different groups of students. Another interesting matter is that the drop-out rates in the current study were quite high (up to 34% for German students in the Dutch track). Therefore, it seems recommendable in general to provide more guidance to students when they are choosing their programme of study, and to continue to investigate what causes students to drop out. Of course, the reasons for students to drop out can be manifold. They will not always be related to students' choice of study programme or their academic achievements, and will therefore not always be preventable.

6.4.2.5 General observations

While we found an L1 advantage on Mean grade and Weighted grade, we found no significant difference between the three groups who studied in an L2 on any of the measures (although, as mentioned above, there was a trend on Drop-out). In other words, the Dutch students in the English track did not differ from the German students in the Dutch or English track. We had expected an advantage for students who study in their home country (i.e., the Dutch students), because they are familiar with the educational system. On the other hand, an advantage for the German over the Dutch students could have been expected as well: Rienties, Beausaert, Grohnert, Niemandsverdriet and Kommers (2012) found that western international students in the Netherlands actually outperformed Dutch students in terms of grades and ECs. They explain this finding with the fact that western international students in general are one or two years older than Dutch students (this was also true in our sample), and that studying abroad is a conscious choice for them. Perhaps in the current study these advantages were cancelled out by the fact that German students who come to the Netherlands to study psychology often have not been offered a spot in Germany because of the highly selective, grade-based admission procedure for psychology in Germany. A priority for future studies should be to get more insight in the academic and linguistic background of the German students who come to the Netherlands.

Something else we would like to comment on is the finding from Chapter 5 that the German students in the Dutch track did not differ significantly from Dutch students in the Dutch track in terms of lexical richness. If these German students indeed were (near-)native in Dutch, as we speculated in Chapter 5, then we would have expected them to not differ in study success from the Dutch students in the Dutch track. The finding that the German students in the Dutch track did obtain lower grades suggests that they were perhaps less proficient in Dutch after all. We only measured lexical richness, but not grammar, errors in vocabulary use, and other aspects of language proficiency. In any of these domains, the German students in the Dutch track may have been less proficient than Dutch native speakers.

Another potential explanation for the finding that the German students in the Dutch track obtained lower grades than the Dutch students in the Dutch track, even if they would have been (near-)native in Dutch, is that the English proficiency of the German students in the Dutch track may have been lower than that of the Dutch students in the Dutch track. If so, they may have struggled more with reading the text books and articles.

Finally, even if the German students were equally proficient as the Dutch students in both Dutch and English, they may have been disadvantaged by factors that were non-language related, such as homesickness, being unfamiliar with the Dutch educational system, etc. More research is needed to evaluate these explanations. To begin with, we should measure various aspects of the Dutch and English proficiency of the students in the Dutch track. Once language proficiency can be adequately taken into account, any remaining effects can be explained in terms of nationality, cultural differences and living abroad.

6.4.3 Does lexical richness affect study success?

6.4.3.1 Grade

The students' grades were not affected by their scores on lexical density, lexical sophistication and lexical variation. Especially the absence of an effect of lexical sophistication is interesting, because in Chapter 5 (see 5.3.3.1) we found a significant relationship between the lexical sophistication of students' answers to open exam questions, and the grades the students had received for these answers (recall that the grades were based on the content, and not on the linguistic quality of an answer). Thus, while there did seem to be a direct relationship between the lexical sophistication of a written answer and its content, this relationship disappeared when considering first-year grades in general.

An explanation may be that first-year students are graded on many different tasks, including multiple-choice questions and quantitative questions, such as statistical computations. Since lexical sophistication is a measure that applies to someone's productive vocabulary, it should not necessarily be associated with students' ability to answer a multiple-choice question or perform a computation. This argument of course also applies to the measures of lexical density and lexical variation, but on those measures no relationship with grades on exam questions was found to begin with (see 5.3.2.1 and 5.3.4.1). In conclusion, there seems to be no relationship between students' productive vocabulary knowledge and the grades they obtain in the first year of study.

6.4.3.2 Drop-out

Considering the above null result on Grade, the finding that lexical density was significantly associated with students' drop-out rates was relatively surprising. The higher the lexical density (i.e., the higher the proportion of content words in students' writing samples), the lower was their chance to drop out of the psychology programme. One should keep in mind, however, that this result was only significant after five outlying and influential cases had been removed from the data set. While the robustness of this measure therefore needs to be

established in further research, it does seem to have potential for identifying students at risk of dropping out.

Fonteyne et al. (2014) developed a model to identify students at risk of not passing the first year of university. Their model could identify 25% of the students who would not pass, with a specificity of 98%. This means that virtually all of the students that were singled out by the model would indeed go on to fail the first year, although the model could only identify a minority of all of the students who would fail. The predictors found to be significant by Fonteyne et al. were academic self-efficacy (i.e., the belief that one will be able to succeed academically, see Fonteyne et al., 2014, p. 347), hours of mathematics instruction in secondary education, results on a basic mathematics test, and vocabulary knowledge. Lexical density may be an interesting predictor to add to such predictive models. These models can be used to offer early interventions and support to students at risk of dropping out. Note that this does not mean that we are arguing that university admission procedures should be based on students' lexical density; at the moment the robustness of this measure is not clear enough.

The exact nature of the potential relationship of lexical density with drop-out rates is not obvious either. The relationship does not seem to be a causal one, where the use of a relatively low number of content words would cause students to drop out. Apparently the explanation also is not that students with lower density scores obtain lower grades and would therefore drop out, since lexical density was not a significant predictor of first-year grades. Perhaps there is a third variable that underlies both someone's lexical density score and his/her chance to drop out. Fonteyne et al. (2014) found that Dutch vocabulary knowledge, as measured with the LexTALE test (Lemhöfer & Broersma, 2012), was a significant predictor of passing the first year of study. The LexTALE test provides an indication of someone's receptive vocabulary. Perhaps the receptive vocabulary is directly related to grades, as well as to lexical density. Many other underlying third variables are conceivable, such as intelligence, memory and attention.

6.5 SUMMARY AND CONCLUSIONS

In the current study, we asked three questions regarding the relationship between students' native language (which we assumed to be directly predictable from their nationality), the language in which they studied, and their study success. To begin with, we showed that students who study in their L1 (here, Dutch students who study in Dutch) outperform students who study in an L2 (here, Dutch students who study in English, or German students who study in Dutch or English) when it comes to grades. In this study, the Dutch students in the Dutch track also obtained more ECs than the other three groups, but not significantly so. With regard to dropping out, the Dutch students (in both tracks) seemed to drop out less often than the German students (in both tracks), but again not significantly so.

These results were obtained after the Dutch students in the Dutch and English tracks had been matched on their English high school grades – in our original sample, the Dutch students in the English track on average had obtained a better English grade in high school

than the Dutch students in the Dutch track. The overall high school grade, however, did not differ between the two Dutch groups in the original sample.

We also investigated whether three measures of productive L2 vocabulary knowledge could predict the students' first-year grades and drop-out rates. These measures were lexical density, lexical sophistication and lexical variation. Higher scores on lexical density (i.e., the ratio of content words to the total number of words in a written text) were associated with lower drop-out rates. All other tests were non-significant.

Offering English-medium education in the Netherlands currently is the topic of fierce debate in Dutch newspapers and media. Among the many voices, hardly any come from empirical research. With the current study, we strove to add empirical, objective insights to the debate. Using data that were provided by the Psychology educational institute at Radboud University, we found evidence that studying in an L2 is disadvantageous when it comes to students' grades. For the Dutch students in our sample, we know that this effect was not caused by pre-existing differences between those students who chose to study in Dutch versus English (as measured in terms of their high school grades). Because the use of English in Dutch education is advancing day by day, it is adamant that we extend our knowledge of the potential effects this may have on students' education, be these effects positive or negative.

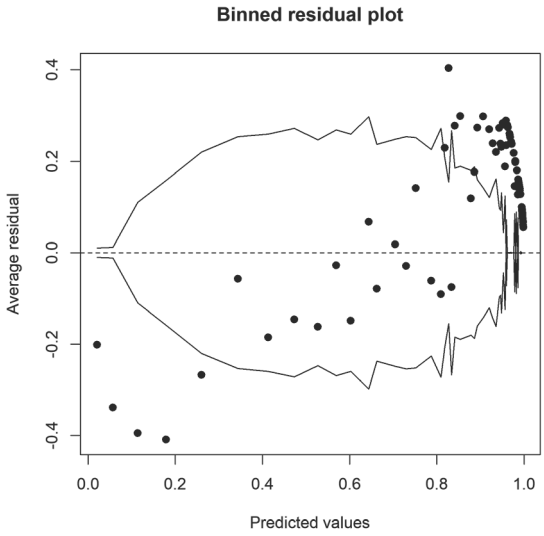
APPENDIX A: MODEL DIAGNOSTICS FOR QUESTION 2

Passing a course/Number of ECs

Since Passing a course was a repeated-measures variable (there were 13 courses), the first approach we tried was to model this variable with a mixed-effects model. Passing a course was a binary variable, therefore we modelled it with a generalised linear mixed-effects model using the logit link function. Random intercepts at the student and course level were included in the model to account for the dependency between the repeated measures and to acknowledge random variation.

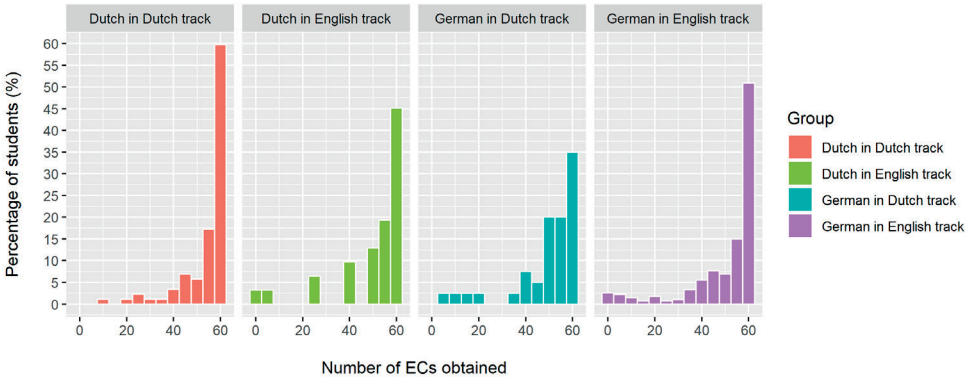
To check whether this model was a good fit to the data, we ran the diagnostics for linear mixed-effects models that were described in the Methods section of Chapter 5 (5.2.4.5), beginning with a residual plot. Because logistic models yield discrete residuals, we created a binned residual plot rather than a regular residual plot, using the *arm* package (version 1.10-1, Gelman & Su, 2018). In binned residual plots, the data are divided into categories (bins) based on their predicted values. Then, for each bin, the average residual is plotted versus the average predicted value for each bin (see Gelman & Hill, 2007, p. 97). Figure A shows the binned residual plot for this model, indicating that the model did not seem to be trustworthy: While no pattern should be visible in the residuals, in this case there clearly was a positive relationship between the predicted values and the residuals. Furthermore, many of the model's predictions fell outside the theoretical 95% error bounds (based on chance, only 5% of the observations should lie outside of these lines). Looking at other diagnostics, we also saw that the subject intercepts were not normally distributed, and neither were the residuals. For these reasons, we decided against using this model.

Figure A. Binned residual plot for the generalised linear mixed-effects model used to model Passing a course from Group.



Next, we considered using ANOVA to compare the average number of ECs that the students in the four groups obtained (please recall that the number of ECs that a student obtained was directly related to how many courses he/she had passed). However, in this case we were confronted with the fact that in all four groups the data were heavily skewed, and therefore non-normally distributed. This is illustrated in Figure B. It means that ANOVA was not an option either, since this technique requires the data in each of the groups to be normally distributed. Although Field et al. (2012, p. 413) write that "when group sizes are equal the *F*-statistic can be quite robust to violations of normality", in our case this was not helpful because group sizes were far from equal.

Figure B. Histograms of Number of ECs obtained, per group.

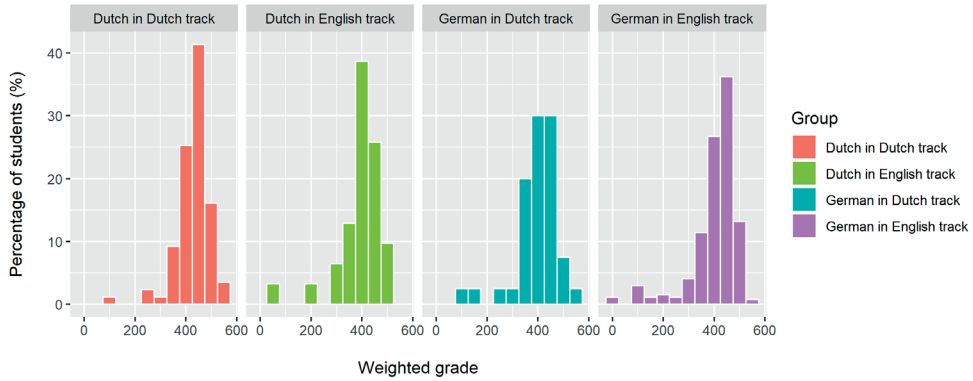


While transformations can sometimes be helpful to restore normality, the Number of ECs variable is especially tricky since 50% of the students had achieved the same (top) score of 60 ECs (i.e., one year's worth of ECs). Thus, even if we did (for example) a log transform, this would not change the fact that the highest score was obtained by about half of the students. The fact that almost all courses (86%) were passed may in fact be the explanation for the low quality of the mixed-effects model of these data. Transforming the data into a binary variable (did someone obtain 60 ECs or not?) in order to do a logistic regression did not seem theoretically justified, since students who scored close to 60 ECs also did very well. At this point, we decided to use a robust ANOVA with bootstrapping to model Number of ECs between the four groups. This is explained in the Methods section of the main text (6.2.4.3). We preferred this approach to the alternative of using a rank-based test such as the Kruskal-Wallis test because "bootstrapping is usually more accurate than traditional approaches" (Wright et al., 2011, p. 254, citing Efron & Tibshirani, 1993), and is expected by Howell (2007, p. 636) to overtake other non-parametric tests (such as rank-based tests) in the future (as paraphrased in Wright et al., 2011).

Grade and Weighted grade

We explored a linear mixed-effects model for Grade (repeated measures), and an ANOVA for its non-repeated equivalent Mean grade. Both these models looked reliable. Since Weighted grade was not a repeated-measures variable, we only explored ANOVA. Again, skew was an issue (see Figure C). Rather than exploring possible transformations, we decided to also use a robust ANOVA with bootstrapping for Weighted grade, following Field and Wilcox's (2017, p. 37) recommendation that "the safest option is always to consider results based on robust methods." To preserve continuity between the analyses we then decided to also use robust ANOVA with bootstrapping for the analysis of Mean grade.

Figure C. Histograms of Weighted grade, per group.



Drop-out

To investigate the relationship between Group and Drop-out, we used a logistic regression model and checked whether the assumptions of this model were met. Because Group was a categorical predictor and it was the only predictor in this model, for all data points within a group the same outcome would be predicted. Thus, in total there were four different possible outcomes, and for model diagnostics, each group represented one case. None of the standardised residuals had absolute values above 1.96. Thus, there seemed to be no outliers. Then, we inspected the presence of influential cases.

We first calculated the leverage per group. This provides an indication of the influence of the observed value over the predicted values (Field et al., 2012, p. 269). According to Field et al. (2012), guidelines state that a case's leverage should not exceed two times the average leverage (Hoaglin and Welsch, 1978), or three times the average leverage (Steven, 2002). Our average leverage was 0.008. The smallest group of participants (Dutch in English track, n = 33) had a leverage value of 0.030, and the next smallest group (German in Dutch track, n = 50) had a leverage value of 0.020. Both these groups exceeded the two-times-the-average cut-off point, and Dutch in English track also exceeded the three-times cut-off point. This means that the outcome values obtained from these two groups influenced the predicted values

more than expected. However, this can be explained by the sample size imbalance (the two overly influential groups had much smaller sample sizes than the two other groups), and therefore does not seem to be of immediate concern.

DFBeta is another measure of influence, reflecting how model parameters change when one data point is excluded from the analysis (Field et al., 2012, p. 270). The values of DFBeta for all observations and parameters were far removed from the critical value of 1 (Field et al., 2012, p. 340), the largest one being 0.13. This also indicates that the above leverage values should not concern us too much. The estimates from a robust logistic regression model were the same as those in Table 10 (considered until two decimal places), indicating that our model was reliable. Thus, we consider the logistic regression model to be suitable for analysing drop-out rates.

APPENDIX B: MODEL DIAGNOSTICS FOR QUESTION 3

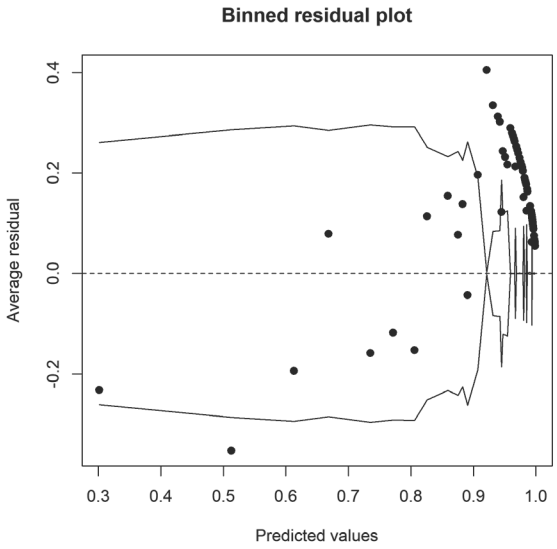
Correlations between the dependent variables

Absence of collinearity (i.e., the absence of high correlations between the independent variables) is an important assumption for all of the models for Question 3. Therefore, we discuss this assumption before zooming in on the specific models. We calculated Pearson's correlation coefficients between the three lexical richness variables (based on visual inspection, the lexical richness variables seemed to be normally distributed). The correlation between LD and LV was $r = .06$, $p = .33$, and the correlation between LS and LV was $r = .01$, $p = .86$. Only LD and LS were significantly correlated: $r = .38$, $p < .001$. Field et al. (2012, p. 276) state that correlations of $r > .80$ or .90 are a cause for concern, so our highest correlation of $r = .38$ does not seem problematic.

Passing a course/Number of ECs

Modelling the Passing a course/Number of ECs variable was problematic, like it was in Question 2 (see Appendix A). First we ran a series of generalised linear mixed-effects models that predicted Passing a course from Group, LD, LS and LV, with random intercepts at the subject and course level. The binned residual plot for the full model that contained all predictors (Figure D) showed that the model was not reliable: Most of the data points fell outside the theoretical 95% error bounds, and there seemed to be a positive trend in the data as well. Therefore, we decided not to use this model.

Figure D. Binned residual plot for the generalised linear mixed-effects model used to model Passing a course from Group, LD, LS and LV.



A regression model (whether robust or not) predicting the Number of ECs from Group and the lexical richness variables was not a good option either because many students had obtained the same (maximum) score of 60 ECs. On the other hand, the lexical richness scores were quite spread out. This suggests that a possible relationship between number of ECs and lexical richness may not have been linear. Due to these issues, we decided to not include Passing a course/Number of ECs in the analysis of Question 3.

Grade

For Grade, we ran the same generalised linear mixed-effects models as for Passing a course. Thus, grades were predicted from Group, LD, LS and LV, with random intercepts at the subject and course level. For the full model that included all predictors, the residual plot (Figure E) showed no heteroscedasticity or patterns, except for the stripes. They are simply caused by the fact that final grades at Radboud University are not continuous, but rather integers (e.g., 6) or half-integers (e.g., 6.5). The grade of 5.5 is never assigned as a final grade, which explains the one stripe that seems to be missing. This is not a problem. The model's residuals also were normally distributed (Figure F).

Figure E. Residual versus predicted values for Grade, as predicted from Group, LD, LS and LV.

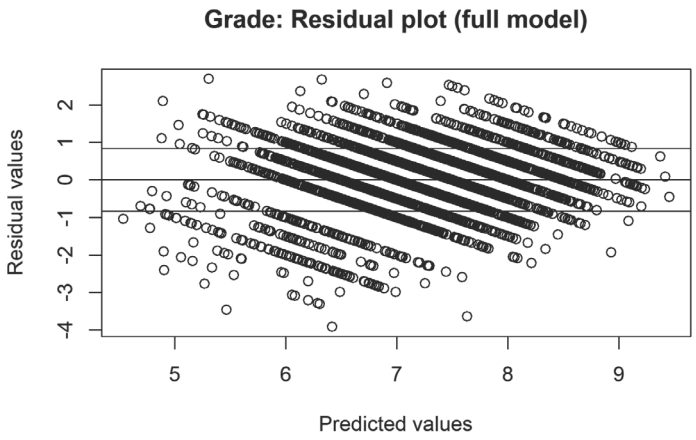
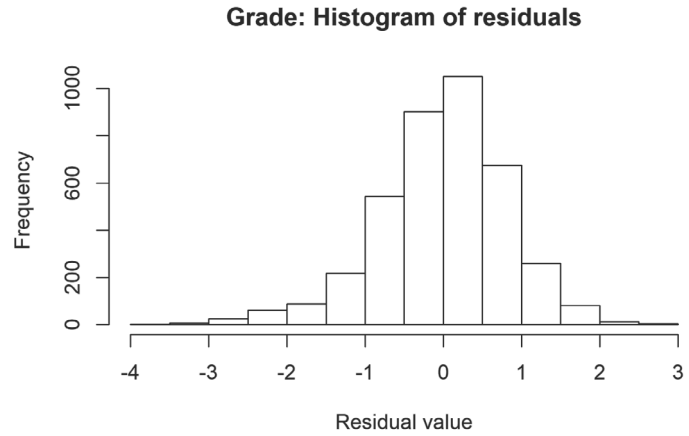
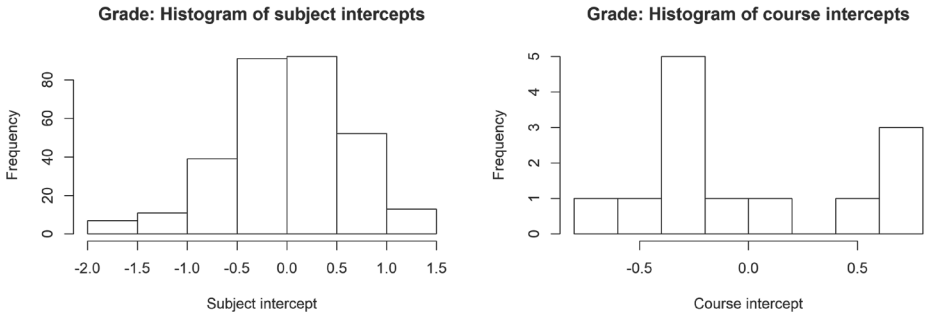


Figure F. Histogram of the residuals for Grade, as predicted from Group, LD, LS and LV.



Cook's distance values showed there were no influential cases, the highest value being .07 (following McDonald, 2002, we only considered values > 0.85 reason for concern). The random intercepts at the subject level seemed to be normally distributed based on visual inspection (see Figure G). The random intercepts at the course level did not seem to be (Figure G), but since there were only 13 data points (because there were 13 first-year courses), we were not very concerned about this.

Figure G. Histograms of the subject and course intercepts for Grade, as predicted from Group, LD, LS and LV.



Weighted grade

In contrast to Grade, the Weighted grade variable should not be used as repeated measures, but only as a sum (for conceptual reasons that are explained in the Methods section, 6.2.3.1). This means that if we first want to examine the effect of Group, and then examine the added value of the three lexical richness variables, we should run a hierarchical regression. However, the first model in this hierarchical regression, which only includes Group as a predictor, would technically be the same model as the ANOVA described in Appendix A. There, we concluded that such an ANOVA would not be reliable because of the non-normal distribution of Weighted grade within each of the groups. This issue, combined with the fact that the Weighted grade variable had not seemed to be of added value in the analysis of Question 2, led us to decide to not include Weighted grade in the analysis of Question 3.

Drop-out

For Drop-out, we ran a forced entry logistic regression with Group, LD, LS and LV as predictors, after having concluded that a hierarchical logistic regression was not feasible (see section 6.2.4.5). In the full model, 8.5% of the residuals had absolute values above 1.96, while according to chance, no more than 5% of the residuals should be of this magnitude. None of the residuals had absolute values above 2.58 (according to chance, this should be around 1%). Forty-eight data points had a leverage value that was more than two times the average leverage, which is problematic (see Appendix A). Four data points had a DFBeta value over the critical value of 1 (range: 1.01–1.37). We followed the approach outlined in Field et al. (2012) for dealing with outliers and influential cases. This means we identified those cases which were both an outlier (i.e., had a standardised residual > |1.96|), and which were influential (i.e., whose leverage exceeded twice the average, or whose DFBeta exceeded 1). There were five of these cases. As mentioned in the Methods section of the main text (6.2.4.5), we ran the model twice, once including the five influential and outlying cases, and once excluding them.

7.
General discussion

This thesis aimed to improve our understanding of naturalistic L2 word learning and lexical development. In the General introduction, I explained that what is considered the essence of naturalistic learning varies between researchers. Some emphasise the location where learning takes place (within the target language community, or outside of the L2 classroom), and others emphasise the characteristics of the setting (it should be informal and unstructured, or language-related tuition should be absent). It is usually also considered important that the communication learners engage in is meaning-driven. In order to do justice to this diversity of definitions, all of these aspects of naturalistic learning were covered in at least one of our studies. I will first provide a short summary of the results of each study, and then consider which insights can be extracted from this thesis as a whole. They fall into three categories: new insights into naturalistic L2 word learning, methodological innovations in lab-based word learning studies, and an empirical contribution to the debate on English-medium instruction in Dutch higher education.

7.1 OVERVIEW OF FINDINGS PER CHAPTER

Chapter 2 was a meta-analysis and meta-regression of 32 studies on incidental L2 word learning from spoken input. In all included studies, the learning was also naturalistic in the sense that the learning contexts were meaning-driven rather than explicitly focused on word learning. The meta-analysis showed that word learning in meaning-focused activities is very well possible: The average learning effect over all studies was large. In the meta-regression, we added five predictors to our L2 word learning model. We found that highly-educated adults were better word learners than children, and that interactive tasks led to more learning than tasks in which the learners only heard someone else speak, but did not interact with that person. The other two task types we investigated, namely audio tasks (e.g., listening to an audiobook) and audiovisual tasks (e.g., watching a film) were not involved in any significant contrasts. The learners scored better on recognition tests than on recall tests. Finally, we evaluated two methodological predictors that concerned how researchers deal with participants' pre-existing knowledge of the target words. We saw that studies which included a no-input control group yielded smaller effects than studies without such a control group: The presence of a no-input control group apparently prevents an overestimation of the learning effect. However, we found no difference in effect size magnitudes between studies in which learning was calculated by means of pre-test to post-test gain scores, as compared to post-tests only.

In **Chapter 3**, we developed and tested an experimental paradigm for studying naturalistic L2 word learning in the lab. We found that it was possible to keep the participants unaware of the fact that they were taking part in a word learning experiment. This enabled us to study naturalistic learning in a highly controlled experimental setting. Still, the learning was meaning-driven, and took place without language-related instruction, in the target language community and outside of the L2 classroom. In an interactive task, the participants were able

to learn many new words. Cognate words were easier for the participants to learn than non-cognates, and there was more learning after four than after two repetitions. Whether three or seven trials appeared between the trial in which a participant was exposed to a word and the trial in which he/she was tested on it did not affect the participants' scores. Twenty minutes after the experiment the participants had forgotten about 24% of what they had learned, and six months after the experiment about 68%. Still, even after all this time they could still remember a substantial and significant amount.

For the study described in **Chapter 4** we used the same paradigm as in Chapter 3, but this time to evaluate an important theory in the field of second language acquisition, namely Swain's Output Hypothesis (1995). We found support for the idea that making learners aware of holes in their L2 vocabulary assists the uptake of these unknown words from the language input. Fifteen minutes after the first post-test, the newly acquired word knowledge had declined slightly, but not significantly. Pre-existing passive knowledge of the target words facilitated learning in those participants who had noticed holes, but actually hindered learning in those participants who had not.

In **Chapter 5**, we tracked the Dutch and English lexical development of Dutch and German psychology students during their first year at Radboud University (Nijmegen, the Netherlands) with their degree being taught either in Dutch or in English. The language learning in this study was naturalistic in all senses, except that English was not the community language outside of the university. Lexical development was operationalised as lexical richness, consisting of the dimensions of lexical density, lexical sophistication and lexical variation. We found that the development of lexical richness in the study language did not differ between Dutch students who studied in L1 Dutch or L2 English. Similarly, lexical richness did not develop differently between German students who studied in L2 Dutch (the community language) or in L2 English. We did find that Dutch students' lexical density and lexical sophistication scores were higher in written Dutch as compared to Dutch or German students' scores in written English. However, it is unknown whether this was an effect of nativeness or a main effect of language. Due to the characteristics of the data set, it was also not possible to draw conclusions about the absolute development of students' lexical richness throughout the year.

Finally, **Chapter 6** showed that students who study in their native country and language (i.e., Dutch students in the Dutch track) outperformed other student groups (i.e., Dutch students in the English track, and German students in either track) in terms of grades. The Dutch students in the Dutch track also obtained more European Credits (ECs), but not significantly so. Dutch students in either track seemed to drop out less often than German students in either track, but again not significantly so. These results were obtained after the Dutch students in the two tracks had been matched on their English high school grades. This was necessary because

the Dutch students in the English track already had better English skills before commencing their studies than the Dutch students in the Dutch track, but even in the Dutch track the students would encounter some English study materials. Students' lexical richness in their study language did not affect their grades, although students who scored better on lexical density dropped out significantly less often.

7.2. NEW INSIGHTS INTO NATURALISTIC L2 WORD LEARNING

7.2.1 A large effect of naturalistic L2 word learning

In Chapters 2 and 3 we investigated how much vocabulary L2 learners can acquire from incidental/naturalistic exposure to input. Chapter 2 was a meta-analysis of existing incidental learning research, where the main criterion for inclusion in the analysis was that the activities the learners engaged in were meaning-driven. In Chapter 3, the learning also was meaning-driven, and in addition took place without language-related tuition, in the target language community and outside of the L2 classroom. Therefore, we considered the learning in Chapter 3 to be naturalistic. Especially in Chapter 3, there seems to have been a large element of intention in the participants' learning behaviour, which was probably due to the design of our interactive learning task. Chapter 2 encompassed 32 different studies, also most likely with different degrees of intention in the learners (please recall that we did not define incidental learning as necessarily taking place without intention, because learners' intentions cannot be measured or controlled). All in all, we conclude that naturalistic exposure to L2 vocabulary, in combination with (some) intention to learn words on the side of the learners, can result in large learning effects. To the general public, this outcome probably would come as no surprise: Anecdotally, there seems to be an understanding that naturalistic L2 exposure is what makes you 'really learn' another language.

7.2.2 Six variables that influence naturalistic L2 word learning

We identified six variables that influence the number of words that learners are able to pick up from naturalistic L2 input. The meta-regression showed a medium-sized effect of age, or more precisely, a combined age/education effect: University students learned more words than children. The exact contribution of age and educational attainment still needs to be untangled in future research, and it is adamant that adults other than university students will be included in the samples. The meta-regression also showed a small-to-medium effect of the treatment the learners engaged in: Interactive, meaning-focused tasks led to higher rates of vocabulary acquisition than non-interactive meaning-focused tasks. There were no significant differences between meaning-focused tasks (either with or without interaction) and situations in which learners only listened to input (e.g., to an audiobook) or were exposed to audiovisual input (e.g., watching a film). However, the magnitude of the audio and audiovisual versus interactive task contrasts were still of medium magnitude. This indicates that it is worthwhile to further investigate these contrasts in a larger sample.

Furthermore, in our first lab study (Chapter 3), we found that cognate status exerted a large effect on the learnability of words: Cognates were acquired at higher rates than non-cognates. An exposure frequency of four exposures to the target words as compared to two exposures also benefited the participants' vocabulary acquisition, with a large effect size. In the second lab study (Chapter 4), we found that noticing the hole was another variable which benefits word learning: When learners become aware that they have a hole in their vocabulary knowledge, they are more likely to learn the word in question from later input. The effect size was medium-to-large.

Noticing the hole interacted with the participants' pre-existing passive knowledge of the target words: Passive knowledge of words that someone had not yet mastered actively was found to facilitate word learning in participants who had noticed holes in their vocabulary. However, participants who had not noticed holes in their vocabulary actually scored higher on words they had had no passive knowledge of prior to the experiment. At a first glance, this finding seems very curious, but we explained the latter result by arguing that people pay more attention to novel stimuli (i.e., to words of which they had no pre-existing passive knowledge). The participants who had become aware of vocabulary holes (i.e., of gaps in their active vocabulary knowledge) would have also paid attention to the words they already had passive knowledge of, because they were prompted to produce these words, but failed. Therefore, the participants' acquisition of these words would have been facilitated by their pre-existing passive knowledge.

7.2.3 Potential mechanisms behind naturalistic L2 word learning

As a third insight, we not only strived to identify predictors, but also potential mechanisms behind naturalistic L2 word learning. Specifically, we considered the potential mechanism behind the effect of noticing the hole in Chapter 4. We hypothesised that noticing the hole may trigger curiosity in learners as to what the missing word form may be, which in turn would lead them to pay more attention to the input. A relationship between attention and L2 word learning has been proposed on the basis of theoretical arguments (e.g., Schmidt, 2001), and has also been established empirically (e.g., Godfroid, Boers & Housen, 2013; Godfroid et al., 2018).

Mediation analysis seems the appropriate technique to statistically investigate this potential relationship between noticing the hole, curiosity, attention and word learning. However, in our data set it was not possible to run such an analysis because the participants in the control group were reassigned to two subgroups based on their self-induced noticing behaviour. As a result, statistically speaking it would be impossible to determine whether a third variable, say attention, would have been a mediating or a confounding variable. Was (say) attention the missing link between noticing the hole and word learning (i.e., was it a mediating variable), or did attention both cause some control participants, but not others, to notice holes in their vocabulary, and facilitate learning (i.e., was it a confounding variable)? Therefore, in Chapter 4 we could draw no conclusions about potential mechanisms behind

naturalistic L2 word learning, but we made recommendations as to how this could be studied in future research.

7.2.4 The development of naturalistically acquired L2 word knowledge

The fourth insight from this thesis concerns how naturalistically acquired L2 word knowledge develops. To begin with, and as was entirely expected, the meta-regression on meaning-focused learning showed that passive knowledge of new words, as measured through recognition tests, was acquired more easily than active knowledge, as measured through recall tests. The effect size was small. While the recognition > recall pattern had been hypothesised before (Nation, 2001), as well as empirically shown (e.g., Brown, Waring & Donkaewbua, 2008; Ellis & He, 1999; Van Zeeland & Schmitt, 2013), it had not yet been quantified for meaning-focused L2 word learning, generalised over a variety of different incidental learning contexts.

Another aspect of vocabulary learning is forgetting over time. On a very small timescale, we found no significant difference between testing the participants three versus seven trials after they had last been exposed to a target word. This was approximately a difference of 20 seconds. Thus, the first lab study (Chapter 3) showed that newly acquired L2 word knowledge does not decay at such a fast rate. In the second lab study (Chapter 4), we extended the time window under investigation, and considered the results of two post-tests that were approximately 15 minutes apart. We found no significant difference between these two post-tests, although there was a small trend in the direction of decay.

In the first lab study, we also contrasted scores on a post-test that was conducted about 20 minutes after the end of the learning phase to scores that the participants obtained during the learning phase itself (i.e., during the price comparison task). Because the price comparison task took approximately 40 minutes and the target words were spread out evenly over the task, the interval between the two testing moments varied between 20-60 minutes. This time, we did detect a decay in word knowledge, with a large effect size. However, this comparison is different from the two above contrasts because here the two tests did not have the same format. In the price comparison task, we measured short-term word recall less than a minute after participants had been exposed to a word, whereas the post-test measured long-term recall after 20-60 minutes through a picture-naming task. This likely explains why we found a large effect, while such an effect had not been detected in Chapter 4 between two post-tests that were 15 minutes apart.

In addition to the first post-test conducted 20 minutes after the end of the learning phase, we also conducted a second picture-naming post-test six months after the learning phase. The forgetting effect size between this second post-test (long-term recall) and the outcomes during the learning phase (short-term recall) was very large. Interestingly, what the participants could remember at this point still was significantly more than 0, although they had only been exposed to the target words four times, six months earlier (or five times in case of an incorrect answer during the first post-test, when the experimenter provided them with the correct word form).

The fact that some L2 words were already forgotten within 20-60 minutes after the experiment, while others could still be recalled after six months, shows that some words seem to be encoded and/or retrieved very differently than others. For example, De Groot and Keijzer (2000) have shown that cognates are less susceptible to forgetting than non-cognates. We also explored the relationship between cognate status and forgetting in an extra statistical model in the appendix of Chapter 3, and found that non-cognates were forgotten at higher rates over a six-month period (but not over a 20-minute period). This exploratory model also showed that for forgetting, it did not matter whether a word had first been recalled three or seven trials after exposure during the learning phase. It would be interesting to expand our knowledge of what makes naturalistically acquired L2 words more likely to be forgotten. If such factors are known, then these words should perhaps receive more attention in L2 text books and classrooms, since they cannot be expected to be retained easily in a naturalistic setting.

7.2.5 Factors that influence effect size magnitudes

The fifth and final insight does not concern naturalistic L2 word knowledge directly, but rather the way it is measured in experimental research. In the meta-regression, we investigated the magnitude of estimated learning effects in an experimental group that was involved in some kind of meaning-driven L2 word learning activity. Specifically, we investigated whether the magnitude of such estimates changed when studies also included a no-input control group, and when they calculated learning scores by comparing pre-test to post-test gains.

Regarding the inclusion of a no-input control group, we indeed found that this was associated with smaller effect sizes, as we had expected (the magnitude of this effect was small). Even using a pre-test to measure participants' pre-existing knowledge does not take away the added value of a control group. In Chapter 3, we conducted a pre-test in order to select (productively) unknown words for each participant. Nevertheless, during the following price comparison task the participants in the control group still managed to produce about 8.3% of phonemes in cognates correctly, and 1.5% in non-cognates (without any exposure to input). This shows that the pre-test apparently had not detected all of their existing knowledge. This underlines the importance of including control groups in studies that work with natural language and aim at estimating learning effects in an absolute sense. In studies that aim to compare L2 word learning in two or more treatment groups (e.g., the lab study in Chapter 4), the presence of a no-input control group is less necessary, although it is never disadvantageous and would facilitate the estimation of absolute learning effects in future meta-analyses.

Unexpectedly, in the meta-regression we found no effect of the use of gain scores. I should note that in the studies that did not make use of gain scores as a way of controlling for participants' pre-existing knowledge, this knowledge was always controlled in another way (for example by means of a control group, or a very careful selection of the target words).

For this reason, we had expected that the use of gain scores would actually inflate learning effect size magnitudes, because the pre-test can draw participants' attention to the target words and invoke the expectation of a later post-test. Both can be expected to lead to more word learning. The fact that we did not find an effect of gain score inclusion (the effect size being almost 0) indicates that it is possible to use pre-test to post-test gain scores without unwanted side effects. Of course, researchers should take very good care of the way in which they present the pre-test to the participants, in order for it not to attract too much attention.

7.3 METHODOLOGICAL INNOVATIONS IN THE LAB STUDIES

In the General introduction (Chapter 1), I mentioned the tension between experimental control and naturalness in naturalistic learning studies. In the two experimental studies in this thesis (Chapters 3 and 4), we strived to maintain experimental control without it coming at the cost of naturalness. In doing so, we focused on two points, namely keeping the participants unaware of our studies' language learning purpose, and using natural language items, by selecting the target items for each participant on an individual basis. Furthermore, we developed a new method for scoring and analysing learners' productive word knowledge that is more sensitive than simply scoring word productions as correct or incorrect. In this section, I will consider the implementation and impact of these three innovations.

7.3.1 Keeping the participants unaware

In Chapter 3 we discussed how participants in lab studies can typically deduce that a study must have something to do with language learning, even if the learning is supposed to be incidental. For example, participants sometimes are recruited based on their L1 or L2 skills, or during the experiment encounter a language which is not their native one. It seems likely that when participants know, or suspect, that they are taking part in a language-related experiment, the learning will be less naturalistic than when participants are unaware of this.

Therefore, we strived to keep our participants unaware of the language learning aspect in our two experimental studies. We did so by telling them an elaborate cover story about our experiment concerning object prices. In addition, we did not tell the participants, who were all native speakers of German, that they had been selected by our participant recruitment system based on their language background. Since the studies were conducted in the Netherlands, it was not suspicious to the participants that the language of communication during the experiment was Dutch. These approaches to hiding the studies' language learning purpose worked very well: Out of the total of 126 participants (in both studies), only one correctly guessed that the study had been about L2 word learning. She was excluded from the analysis.

The idea that keeping participants unaware of a study's language learning purpose is important in order to approximate naturalistic L2 learning in the lab was also adopted by Koch, De Vos, Lemhöfer, Housen and Godfroid (2019). That study concerned L1 Dutch speakers' acquisition of the stem vowel change in L2 German verb conjugation. For example, in the

third person singular, the verb *geben* (English: *give*) takes the form *er gibt* (English: *he gives*). As in my studies, the participants were exposed to the phenomenon under investigation in a meaning-driven task. This time, they saw several pictures on a computer screen and had to form a sentence that included these pictures. Like in Chapter 3, this involved an alternation between uttering such a sentence, and listening to an L1 experimenter uttering these sentences. Both in the experimenter's and the participant's trials, the to-be-used verb sometimes (but not always) required a vowel change.

Koch et al.'s (2019) cover story was that the experiment would be conducted in a variety of languages to investigate how the language we use influences the way we think. In reality, it was only conducted in German. There were two conditions. The participants in the explicit condition were told the cover story, but also received information on the phenomenon of the German stem vowel change. They were told they should pay attention to how the experimenter produced these verbs, and they should try to produce them correctly. In the incidental condition, the participants were only told the cover story. Later interviews with the participants revealed that of the 28 participants in the incidental condition, only one had guessed correctly that the experiment was about the German stem vowel change. She was excluded from the analysis. The remaining 27 participants had all believed the experiment was about how language influences our thoughts. Still, of those 27 participants, 21 indicated that they had noticed the stem vowel changes in the experimenter's utterances, and six had not.

The learning scores of these 21 participants in the incidental condition were compared with those of the 21 participants in the explicit condition, all of whom had received explicit instruction about stem vowel changes, and therefore had also noticed them. During the experiment, the participants had been prompted to produce some German verb forms before having received the correct input from the experimenter. These productions functioned as a pre-test of the participants' ability to correctly conjugate verbs that require a stem vowel change. On those trials, the participants in the explicit condition outperformed those in the incidental condition, with a large effect size. This suggests that explicit instruction affects the correctness of the learners' German morphosyntax. However, the improvement due to input from the experimenter was equally large in both conditions. Thus, receiving explicit instruction about the target structure did not result in more learning from input, as compared to noticing the target structure without having received explicit instruction.

It should be noted that even in the explicit condition, the participants had not been told that the primary aim of the experiment was to investigate learners' acquisition of the German stem vowel change. The participants still thought that the study was about the influence of the language that we speak on how we see the world. The null-effect of morphosyntax instruction on the magnitude of the learning effect nevertheless is interesting. Koch et al. (2019) explain this finding referring to the cognitive demands of the meaning-focused task. For example, in the interviews the participants indicated that they also spent a lot of cognitive effort on choosing picture combinations and on case marking, meaning that less

cognitive resources may have been left for focusing on the stem vowel change. In addition, despite having received no instruction on the German stem vowel change, 21 learners in the incidental condition did indicate that they had noticed and paid attention to this phenomenon, which also is likely to also have resulted in learning. It would have been very interesting to also analyse the data of those participants in the incidental condition who had not noticed and paid attention to stem vowel changes, but with $n = 6$ their number was too low. On a side note, the fact that 21 out of 28 participants in the incidental condition reported to have noticed and paid attention to the target structure again shows that even in studies that are designed to target incidental learning, participants may independently develop an intention to learn.

In parallel with the coming about of my PhD thesis, Brandt, Schriefers and Lemhöfer (2019) investigated the naturalistic acquisition of L2 Dutch definite articles, which reflect the grammatical gender of the noun they accompany. Brandt et al. worked with a similar population of German students as I did, who were kept unaware of the study's true aims. The researchers compared two different cover stories, one of which was more effective in hiding the study's aims than the other. During debriefing, out of the 32 participants who had received the first cover story, 19 mentioned that they had paid explicit attention to definite articles in the experimenter's input, and/or suspected that this was what the study was about. For the second cover story, nine out of the 32 participants voiced similar thoughts.

Brandt et al. (2019) found no main effect of cover story (the first versus the second) on the correctness of the participants' use of Dutch definite articles, and also no difference between the groups with regard to how much they learned from the experimenter's input. Brandt et al. then rearranged the data, grouping together the 28 participants who had paid attention to definite articles and/or thought that this was what the study was about, and the 36 participants who had not. With this division, they found that the aware group overall did better than the unaware group in producing Dutch definite articles, and that there was a descriptive (though not a significant) trend towards more learning from the input as well (with a *p*-value of .07).

In short, both Koch et al. (2019) and Brandt et al. (2019) found that awareness of the target structure did not significantly increase learning from input. On the one hand, this suggests that our precautions in hiding the lab studies' aims may have been superfluous. On the other hand, awareness of a target structure is not the same as awareness of a study's aim. Especially in Koch et al., it was still the case that the participants did not expect to be scrutinised on their production of the target structure, as became apparent from interviews (in Brandt et al., no distinction was made in the interviews between the participants' noticing of the target structure, and their potential suspicion that this was what the experiment was about). Therefore, we do not know whether awareness of a study's aim, as opposed to awareness of its target structure, could still have an effect. In addition, Koch et al. and Brandt et al. focused on morphosyntax, the acquisition of which may be different from vocabulary acquisition.

This question of whether awareness of a study's aim influences word learning was addressed directly in a series of studies by Peters and colleagues. In all these studies, the researchers told some participants that they would take a vocabulary post-test after having read a text or having watched a film. Other participants did not receive this information, and were told that they would answer content questions afterwards. For the most part, the researchers found no significant effect of test announcement on L2 word learning (Montero Perez, Peters & Desmet, 2018; Peters, 2007a, 2007b; Sercu, Dewachter, Peters, Kuiken & Vedder, 2006). Peters, Hulstijn, Sercu and Lutjeharms (2009) did find that test announcement significantly benefited word form recognition after reading (i.e., was this word present in the text you just read?), with a small effect size. However, in the same study there was no significant effect on two recall tests (meaning translation with and without context). In Montero Perez, Peters and Desmet (2015), it was the other way around: Post-test announcement significantly benefited participants' scores on a meaning recall test (translating into the L1) with a small-to-medium effect size, but not on a form recognition test.

In summary, the findings from these experimental studies on the effect of vocabulary post-test announcements are mixed. In most cases, there was no effect, but there is also evidence for beneficial effects of post-test announcement on both word form recognition and on meaning recall. In our lab studies, we measured recall of L2 word forms. It seems such a measure was only used by Sercu et al. (2006), who tested "guided productive mastery of the target words" (p. 62), and they found no significant effect of test announcement. Therefore, the evidence on the importance of hiding a study's L2 word learning aspect from participants is still too scarce to draw a conclusion about this central part of experimental design. Until it can conclusively be shown that participants' expectations about the researchers' aims and about the presence of a vocabulary post-test do not affect their word learning behaviour, our paradigm offers a way of experimentally investigating naturalistic L2 word learning without participants being aware of this aim.

7.3.2 Selecting target words for each participant

Another innovation in our first lab study (Chapter 3) was that we selected the to-be-learned target words on an individual basis for each participant during the experimental session. We did so by making the participants produce all the target words in the context of a price judgment task before the participants were exposed to input, and simultaneously coding the correctness of their utterances. When the participants took a break before commencing the next part of the experiment, the software I developed made a personalised selection of unknown target words and known filler words for each participant, based on which words they had and had not been able to produce during the pre-test. In doing so, the program took the words' cognate status, length, L1 frequency and compound status into account.

This procedure allowed us to work with natural language items rather than artificially created words, while ensuring that all participants were exposed to an equal number of (productively) unknown words, and thus experienced a similar cognitive load. This

represents an improvement to existing approaches for including natural language vocabulary in word learning studies. I will now discuss a few examples of such approaches, and their disadvantages. The first approach is to base the selection of target words on a group-level pre-test. For example, Ellis, Tanaka and Yamazaki (1994) let their participants translate 65 L2 English words into L1 Japanese. This was done one month before the rest of the experiment was conducted. Based on the outcomes, Ellis et al. selected 18 target words that were unknown to the participants. While in some ways this is a good way to survey the participants' pre-existing knowledge of the target words and select items accordingly, the disadvantage is that the participants are familiarised with the target words. As a result, they might look them up at home, or they might recognise them during the experiment and pay extra attention to them. Sometimes it is not possible to select target words that are unknown to all participants. For example, Ellis and He (1999) used the same approach, and selected items with an overall non-recognition level of 88%.

An often used alternative to pre-testing in advance of the experiment and basing the item selection on the outcomes of this pre-test, is simply to conduct a pre-test directly before the start of the treatment and calculate pre-test to post-test gain scores. This removes the problem that participants might look up the target words at home, and their pre-existing knowledge can be taken into account in the analyses. However, in Chapter 2 we discussed some issues that are associated with the use of pre-tests. These were that they might raise the expectation of a post-test, and they may highlight the target words to the participants, causing the participants to pay more attention to these words when they later appear in the input. In addition, if the participants' pre-existing knowledge differs widely, some participants will have (many) more new words to learn than others.

It is also possible to include a control group in the experiment, rather than pre-testing the experimental participants on the target words. Either item selection could be based on the pre-test scores of a control group (which should be sampled from the same learner population as the experimental group), or the scores of the experimental group could be compared against the scores of the control group. This removes all problems associated with the use of a pre-test, such as highlighting the target words to the participants. However, it is also less precise: Vocabulary knowledge varies widely, and therefore it cannot be guaranteed that the control group scores are a good estimate of those of the experimental group. Because our metaregression showed that control group inclusion is important in order to not overestimate effect sizes, I still recommend control groups to be included, but not to be relied upon exclusively when it comes to accounting for participants' potential pre-existing knowledge of target items.

Yet another way of taking participants' pre-existing knowledge into account is to ask participants, after the experiment is already completed, to indicate which words they did and did not know already (e.g., Sydorenko, 2010; Winke, Gass & Sydorenko, 2010). These words can then, on an individual basis, be excluded from the analysis. This approach has the advantage that it is only conducted after participants are exposed to input, and thus the target

words have not been highlighted in any way. It does rely on the assumption that participants are able to recognise, in retrospect, what their prior knowledge was like. Sydorenko (2010) included non-words in the post-experiment prior knowledge test to at least control for the participants potentially overestimating their prior knowledge (it is not reported how often the participants claimed that they knew these non-words). One disadvantage of excluding words from the analysis in hindsight is that it is not possible to control how many new words each participant has to learn. Especially when comparing two or more groups, it is desirable that the participants in both groups on average are exposed to an equal number of previously unknown words.

I believe the way of pre-testing and selecting target and filler words which we used in our first lab study (Chapter 3) circumvents all of the issues described above.[1] Because the pre-test was not presented as such, and embedded in the price judgment cover story, our participants did not expect a vocabulary post-test. Furthermore, all of the words that appeared in the experiment, including all fillers, were part of the pre-test. This means that the target words were not highlighted any more than the other words. Because the pre-test was conducted directly before the experiment, the participants could not look them up before the experiment started. Finally, because the selection of target words was tailored to each participants' pre-existing knowledge, this selection did not contain any target words that the participants already had shown any knowledge of, at least not productively. At the same time, because fillers were selected on an individual basis too, we knew for sure that the participants knew all the filler words, and never encountered two unknown words in one trial.

Researchers should always consider what is the best way to control for their participants' pre-existing knowledge. While our approach solves a lot of problems that are associated with the use of a pre-test, alternative approaches may be more desirable based on the design or aims of a particular study. For example, in our second lab study we used the approach that was also used by Sydorenko (2010) and by Winke, Gass & Sydorenko (2010): We asked the participants about their pre-existing word knowledge after they had already completed the experiment, and then excluded the already-known words from the analysis. This was necessary because there was no way in which we could pre-test the participants in the control group without them noticing holes in their vocabulary. However, in this study we based the selection of target and filler items on the pre-test scores from the participants in our first lab study, who came from the same population. The combination of these two approaches should have controlled for the participants' pre-existing knowledge as much as possible given our research question. Because we did have pre-test data for the experimental

[1] For completeness, it should be noted that at least Ellis and Heimbach (1997) also selected target items on an individual basis for each participant after a pre-test (and perhaps there are other studies doing so of which I am not aware). However, the procedure in Ellis and Heimbach (1997) was different from the one in our study, since they manually selected the target items for each participant, and did not seem to hide the goal of the pre-test or of the study itself. However, whether or not this is problematic of course fully depends on the kind of learning that researchers are trying to observe.

participants in the second lab study (the pre-test was what made these participants notice vocabulary holes), we could also evaluate the reliability of the participants' retrospective reports of their pre-existing vocabulary knowledge. When comparing these reports to the participants' actual pre-test scores, we found that they converged for 99.7%. Furthermore, in order to not have too much variation between the participants with regards to the number of words they had to learn, we excluded four participants who indicated that they already had actively known more than 25% of the target words.

7.3.3 Scoring and predicting word knowledge at the phoneme level

In both our two lab studies, we measured how many new words the participants had learned. This was measured as word form recall, where the participants were asked to produce the L2 word forms. We considered the question of how their productions should be scored. Often, the outcome was not binary (i.e., fully correct or fully wrong), but something in between. For example, a participant could produce a word correctly except for one phoneme.

In the literature, researchers have used various approaches to scoring productive word knowledge. The simplest approach still is binary correct/false scoring. Sometimes, researchers do not indicate whether there were any responses that were neither fully correctly nor false, and if so, how they were dealt with (e.g., De la Fuente, 2002; Ellis & He, 1999). Other researchers do explain how they deal with in-between, partially correct answers. For example, Peters (2014, p. 85) also marked words with minor spelling errors as correct, such as *liabilitie* (instead of *liability*). Webb (2005, 2007) was stricter. He investigated Japanese students' acquisition of English orthographic word forms (i.e., of spelling). Only words that were spelled fully correctly were awarded one point, all other productions were awarded 0 points. Webb (2005) motivates this decision by explaining that otherwise it would not be possible to determine whether a student had learned the words through the learning task that he had designed, or was only attempting to write down the word form he/she had just heard aurally.

On the one hand, it seems reasonable that Webb (2005, 2007) marked words with minor mistakes as incorrect, whereas Peters (2014) marked them as correct. This is because Webb was specifically interested in the acquisition of orthography, whereas Peters was investigating form recall in general. She did not provide aural cues, like Webb did. On the other hand, if we borrow the example word from Peters, a student in Webb's study would still have done better when writing *liabilitie* as compared to *leabillithie*. This nuance is not captured in Webb's scoring system, where both productions would have been marked as false. As for Peters' scoring system, how does one determine the boundary between a 'minor' and a 'major' spelling mistake?

Other researchers have developed scoring systems which are more sensitive and leave no room for ambiguity. For example, Nakata (2016) assigned scores of 0.00, 0.25, 0.50, 0.75 and 1.00 based on the number of correctly produced letters. This allows for a more sensitive assessment of the participants' performance, although in this particular study none

of the participants produced any of the words correctly. Still, there seems to be room for improvement: Why did Nakata limit himself to multitudes of 0.25, rather than just counting the proportion of correctly produced letters per word?

Another example of a sensitive and clearly defined scoring system comes from Meara and Ingle (1986), who worked with English learners of French. They tested their learners on target words containing three consonants, and marked the consonants only. Each consonant was scored as correct (1) or incorrect (0), which led to eight (2^3) possible outcomes per word: 111, 110, 101, etc. These eight codings were treated as nominal categories, and the number of responses in each category were compared by means of chi-square tests. This analysis seems suitable for researchers who are interested in comparing different error patterns, but is not as useful when the purpose is to generate one average learning score per participant, or to compare participants' scores between different conditions.

I did not survey the word scoring literature systematically, so the above studies should be taken as an impression of the research. However, my impression was that the large majority of studies uses binary scoring at the word level (thereby potentially failing to detect subtle differences between participants' performances), and uses unclear methods for dealing with word productions that are neither fully correct nor wrong. Like Nagata (2016), we therefore scored the participants' word productions at the phoneme level, but we did not limit ourselves to multitudes of 0.25. To obtain descriptive statistics, we calculated the proportion of correctly produced phonemes.

For inferential statistics, we used the binomial distribution to model the number of correctly produced phonemes per word (for example, four phonemes produced correctly and one incorrectly). This approach is preferred to simply assigning each word a proportion (for example: 0.80), for a number of reasons (Crawley, 2007, pp. 569-570): When using proportions, the errors are not normally distributed, the variance is not constant, the response is bounded (between 0 and 1), and sample size information is lost. As far as I know, we were the first to mark the correctness of word productions at the phoneme level and then analyse those outcomes in linear mixed-effects models based on the binomial distribution. Our approach later was adopted by Mickan, McQueen and Lemhöfer (2019), and Mickan, McQueen, Piai and Lemhöfer (2018, August).

7.4 ENGLISH-MEDIUM INSTRUCTION IN DUTCH HIGHER EDUCATION

Chapters 5 and 6 were motivated by the surge in articles that appeared in the Dutch media regarding the use of English as the language of instruction in Dutch higher education. At the same time, this was also an opportunity to investigate language learning that was naturalistic in practically all senses mentioned in the General introduction. The language learning was informal and unstructured, meaning-driven, and non-instructed. While some learning took place in classrooms, these were not L2 classrooms. Only the last definition, stating that naturalistic language learning should take place in the target language community, only applied to the students in the Dutch track, but not to those in the English track. In this General

discussion I want to focus on the arguments that can be extracted from these two chapters regarding the suitability of using English as the medium of instruction (EMI). I will specifically focus on EMI in the Netherlands.

7.4.1 Strengthening students' language proficiency through the language of instruction

Both proponents and opponents of EMI often refer to the idea that exposure to the language of instruction, or in other words the study language, benefits students' proficiency in that particular language. The difference between the two camps lies in the fact that opponents of EMI stress the importance of Dutch students studying in Dutch in order to continue to improve their (professional) Dutch proficiency. In contrast, those in favour of EMI attach more importance to the development of students' English proficiency, and therefore encourage the use of English in the classroom. Both camps might be surprised to learn that the literature review in Chapter 5 suggested that the benefits of EMI on L2 English language development are either small or non-existent, although the pool of studies is small, and their outcomes often inconclusive. With regard to the hypothesised benefits of Dutch-medium instruction (DMI) on L1 Dutch language development, I did not find any empirical studies. I should point out that with language development resulting from 'exposure to the language of instruction' I mean naturalistic language learning, and not any language learning that might result from courses that are specifically aimed at improving students' language skills (for example, an academic writing course). The latter is outside the scope of this thesis.

I investigated whether the English language proficiency of Dutch students in the English track developed at a different rate from the Dutch language proficiency of Dutch students in the Dutch track. In the Dutch track, all teaching was in Dutch, although some study materials were in English. In the English track, both the instruction and the study materials were in English. The question I investigated is a different one from evaluating whether naturalistic exposure to the study language indeed benefits students' proficiency in that language. While this latter question directly corresponds to the arguments summarised in the paragraph above, it could not be answered with the data set that was available to me. In order to study the absolute effects of study language exposure on proficiency, we would have needed the students to also produce writing samples in the language they did not study in (i.e., English writing samples from students in the Dutch track, and vice versa). Such baseline data would allow the disentanglement of effects that are due to study language exposure in the study context, and other exposure to the study language (for example, watching a film in Dutch or in English at home). In addition, to examine absolute gains over time, the writing samples would need to be comparable in terms of topic and complexity (and perhaps other dimensions too).

Nevertheless, the question of relative differences in study language development also is interesting to the current debate. I detected no difference between the lexical development of Dutch students in the Dutch and English tracks in their respective study language. We had hypothesised to see more improvement in English, simply because there generally is more

room for improvement in an L2 as compared to an L1. This expected advantage for English development over Dutch development was not found, at least not in terms of lexical richness. I nevertheless think that many more prospective Dutch students choose to study in English in order to improve their English language skills, than vice versa for their Dutch language skills. Therefore, in my opinion, the absence of the English development advantage in our data set is an argument in favour of the DMI camp in the current debate. What prospective students can take away from our findings is that they can let their choice of study language be motivated by arguments other than the study language's relative impact on their language skills. An example of another aspect to consider is the language of the job market that students would like to prepare themselves for.

For policy makers who decide upon language use at universities, philosophical and political arguments will continue to play a role in deciding whether Dutch or English language skills are deemed more important for students in the Netherlands. This decision depends on whether one's view of the future is focused more on a student's place in the Netherlands, or as a global citizen. Of course, not everyone finds himself or herself at one of the ends of this spectrum. Many argue that there should be a place for both Dutch and English at Dutch universities (e.g., De Groot, 2017; Maex, 2017; Van Oostendorp, 2017). However, let us not forget that the actual effects of study language exposure on students' language skills are currently still inconclusive, and await further research. In contrast, in Chapter 6 we clearly showed that study language has a significant impact on students' grades. Policy makers (and students) can already let their decisions be guided by this finding, which is discussed in the next section.

7.4.2 Lower grades when Dutch students study in English rather than Dutch

Like students' language development, the quality of education also is an argument that is used to argue both against and in favour of EMI. Those who are in favour state that international classrooms, which are a direct result of using English rather than Dutch as the language of instruction, increase the quality of education (e.g., KNAW, 2017; Lizzini, Martijn, Munk & De Regt, 2017; Maex, 2017; Van Oostendorp, 2017). This is because international students and lecturers would introduce new ideas and perspectives in the classroom. The opponents of EMI do not deny this argument, but instead stress that the quality of education suffers when Dutch students and lecturers have to communicate in an L2 (here, English), rather than in their L1 (here, Dutch) (e.g., Hermans, 2017; Huygen, 2017; Kleinjan, 2017).

Of course, quality of education is a many-sided construct that can be studied and operationalised in many different ways (e.g., students' satisfaction, students' achievements, or accreditation by external examiners). We considered the grades and number of European credits (ECs) that the students in our data set obtained. The exam questions and grading criteria were the same for all the students. Therefore, we assumed that students who received higher grades had better mastered the course's content. A high amount of learning is one outcome of good-quality education. In Chapter 6, we presented the finding that

Dutch students who studied in Dutch obtained significantly higher grades than Dutch students who studied in English. The difference amounted to 0.55 points on a 1-10 scale, a medium-sized effect.

To find the underlying cause of this 0.55 point difference, we investigated if it was related to the students' lexical richness in their study language. This was operationalised through three measures of the students' productive, written vocabulary knowledge. When Dutch students were writing in Dutch as compared to English, they scored higher on two out of three lexical richness measures, namely lexical density and lexical sophistication. On the third measure, lexical variation, there was no difference between the two groups. Nevertheless, none of the lexical richness measures significantly predicted the students' grades.

Trying to explain the grade difference through lexical richness (as we did) is different from trying to explain it by pointing to potential effects of L1 versus L2 classroom communication, which is central to the argumentation of Hermans (2017), Huygen (2017) and Kleinjan (2017), although they do not directly mention students' grades. Following their line of argument, the question arises whether the grade difference could potentially be explained by suboptimal English language use of the Dutch lecturers, the Dutch students, or both. Vinke (1995) already provided some evidence that native Dutch lecturers' command of English is not as good as their command of Dutch. In a study with 16 Dutch lecturers, on average these lecturers were rated less favourably by observers on their variation in intonation and speed of delivery, verbal fluency and the use of vague terms, when the lecturers were teaching in English as compared to Dutch. However, these outcomes were not linked to student outcomes, such as grades. Bouma (2016) also presented a case study in which the English proficiency of three university lecturers was evaluated, but did not link these outcomes to students' achievements either. This is a necessary step to show that lecturers' proficiency in the teaching language would indeed impact the students' results, as it is sometimes claimed in the media.

Although absent for lecturers, there is evidence that students' proficiency in their study language is related to their study success (e.g., De Koning, Loyens, Rikers, Smeets & Van der Molen, 2012, for L1; Fakeye & Ogunsiji, 2009, for L2; Fonteyne, Duyck & De Fruyt, 2014, for L1; Zijlmans, Neijt & Van Hout, 2016, for L2). While we did not find such an effect for lexical richness, there are many ways to operationalise proficiency. Fonteyne et al. (2014) measured study language proficiency with the LexTALE Dutch vocabulary test (Lemhöfer & Broersma, 2012). De Koning et al. (2012) used the Word Matrix section from the Groninger Intelligence Test (Kooreman & Luteijn, 1987), which is a verbal analogy test (e.g., "cow is to calf, as horse is to…"). Zijlmans et al. (2016) used a self-developed test of Dutch as an L2, covering reading, writing, listening and speaking. They used the average score over these four aspects of language proficiency to predict ECs and grades. While the outcomes of these three studies are interesting, none of them specifically focused on students' communicative skills. Although the arguments made by opinion piece writers such as Hermans (2017),

Huygen (2017) and Kleinjan (2017) seem plausible, the empirical evidence to support them is still lacking.

With regard to the 'quality of education' argument, I draw the following conclusions. It should be empirically evaluated whether there is a negative relationship between the language used in the classroom/lecture hall, and the quality of education at Dutch universities. Different operationalisations of education quality should be used, for example (but not exclusively) the students' grades. While we found that that Dutch students obtain lower grades when they study in English rather than Dutch, we could not directly link this result to the state of classroom communication. It could also be due to other factors, such as the students' ability to process or reproduce study material in a non-native language (e.g., Vander Beken & Brysbaert, 2018). Regardless of the underlying causes, our data showed that Dutch students who opt to study in Dutch obtain higher grades than those who opt to study in English. This result was obtained after correcting for pre-existing differences between the groups in terms of high school grades. This is something that Dutch students can take into account when deciding upon the programme in which to enrol. At the same time, our data also showed that they do not need to worry about obtaining less ECs or having a higher chance of dropping out when studying in English.

7.5 CONCLUSION

Naturalistic language learning is multi-faceted and can take place in a wide range of situations, covering both incidental and intentional learning. In this thesis, I presented a collection of studies which covered five different definitions of naturalistic L2 learning. Together, these studies yielded various new insights in the processes and outcomes of naturalistic L2 word learning, such as the finding that (intentional) naturalistic learning can lead to high acquisition rates. Another important contribution of this thesis lies in the methodological domain. I have presented three innovations that are important when conducting research on naturalistic language learning: keeping the participants unaware of a study's language learning purpose, pre-testing them on the materials without this having an impact on the rest of the experiment, and going beyond binary scoring of word utterances. It is my hope that these practices will be adopted and further refined by future researchers.

In the second half of this thesis, I entered in the discussion regarding the use of English as a/the language of instruction in Dutch higher education. Nowadays, more and more university degrees in the Netherlands are offered in English, with the goals of attracting international students and thereby creating international classrooms, as well as to prepare Dutch students for the international job market. I showed that Dutch students who received classroom instruction in Dutch obtained higher grades than Dutch students who were instructed in English. Such an effect was absent regarding their number of obtained ECs and drop-out rates. I also showed that Dutch students had a higher lexical richness when writing in Dutch than in English, but this finding could not be linked to their grades. Notably, higher lexical density scores (i.e., one of the three dimensions of lexical richness) were associated

with lower drop-out rates. Taken together, these findings raise questions about the desirability of using English as the main language of instruction at Dutch universities.


Frederik deVos - 2018

Appendices

REFERENCES

Abraham, L. B. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning, 2*1(3), 199–226. doi:10.1080/09588220802090246

Aguilar, M., & Muñoz, C. (2014). The effect of proficiency on CLIL benefits in Engineering students in Spain. *International Journal of Applied Linguistics, 24*(1), 1–18. doi:10.1111/ijal.12006

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

Aldera, A. S., & Mohsen, M. A. (2013). Annotations in captioned animation: Effects on vocabulary learning and listening skills. *Computers and Education, 68*, 60–75. doi:10.1016/j.compedu.2013.04.018

Al-Homoud, F. (2008). *Vocabulary acquisition via extensive input* (Unpublished doctoral dissertation). University of Nottingham, UK.

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. doi:10.1037//0278-7393.20.5.1063

Baayen, R. H. , Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database* [Database]. Retrieved from http://celex.mpi.nl/

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. doi:10.1037/0882-7974.4.1.3

Baltova, I. (1999). *The effect of subtitles and staged video input on the learning and retention of content and vocabulary in a second language* (Doctoral dissertation). Retrieved from http://hdl.handle.net/1807/13234

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning, 57*(1), 35–56. doi:10.1111/j.1467-9922.2007.00398.x

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. doi:10.18637/jss.v067.i01

Baumgarten, N. (2014). Recurrent multiword sequences in L2 English spoken academic discourse: Developmental perspectives on 1st and 3rd year undergraduate presentational speech. *Nordic Journal of English Studies, 13*(3), 1–32.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A new and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289–300.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. Sebastopol, CA: O'Reilly Media.

Birulés-Muntané, J., & Soto-Faraco, S. (2016). Watching subtitled films can help learning foreign languages. *PLOS ONE, 11(*6), 1–10. doi:10.1371/journal.pone.0158409

Bisson, M.-J., Van Heuven, W. J., Conklin, K., & Tunney, R. J. (2014a). Processing of native and foreign language subtitles in films: An eye tracking study. *Applied Psycholinguistics, 35*(2), 399–418. doi:10.1017/S0142716412000434

Bisson, M.-J., Van Heuven, W. J., Conklin, K., & Tunney, R. J. (2014b). The role of repeated exposure to multimodal input in incidental acquisition of foreign language vocabulary. *Language Learning, 64*(4), 855–877. doi:10.1111/lang.12085

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. doi:10.1017/s1366728914000078

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.) (pp. 221–235). New York, NY: Russell Sage Foundation.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons. doi:10.1002/9780470743386

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning, 67*(2), 348–393. doi:10.1111/lang.12224

Bouma, K. (2016). *Geklaag over gebrekkig Engels in collegezaal neemt toe*. Retrieved from https://www.volkskrant.nl/nieuws-achtergrond/geklaag-over-gebrekkig-engels-in-collegezaal-neemt-toe~bf244322/

Brandt, A. C., Schriefers, H., & Lemhöfer, K. (2019). *The processing of natural corrective syntactic input in second language learners: A laboratory study of language learning in dialogue*. Manuscript in preparation.

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.

Burnage, G. (1990). *CELEX: A guide for users*. Nijmegen: SSN.

Burnham, J. C. (1990). The evolution of editorial peer review. *Journal of the American Medical Association, 263*(10), 1323–1329. doi:10.1001/jama.1990.03440100023003

Carey, S. & Bartlett, E. (1978). Acquiring a single new word. *Proceedings of the Stanford Child Language Conference, 15*, 17–29.

Carroll, J. B. (1967). Foreign language proficiency levels attained by language majors near graduation from college. *Foreign Language Annals, 1*(2), 131–151. doi:10.1111/j.1944-9720.1967.tb00127.x

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

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics – Simulation and Computation, 39*(4), 860–864. doi:10.1080/03610911003650383

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardised assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. doi:10.1037/1040-3590.6.4.284

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum. doi:10.4324/9780203771587

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics, 19*(1), 15–18. doi:10.2307/1268249

Cook R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York, NY: Chapman and Hall.

Craik, F. I. M., & Bialystok, E. (2006). Cognition through the lifespan: Mechanisms of change. *Trends in Cognitive Sciences, 10*(3), 131–138. doi:10.1016/j.tics.2006.01.007

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671–684. doi:10.1016/s0022-5371(72)80001-x

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268–294. doi:10.1037/0096-3445.104.3.268

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

Dafouz, E., Camacho, M., & Urquia, E. (2014). 'Surely they can't do as well': A comparison of business students' academic performance in English-medium and Spanish-as-first-language-medium programmes. *Language and Education, 28*(3), 223–236. doi:10.1080/09500782.2013.808661

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 Graaff, R., & Housen, A. (2009). Investigating the effects and effectiveness of L2 instruction. In M. Long & C. Doughty (Eds.), *The handbook of language teaching* (pp. 726–755). Oxford, UK: Blackwell. doi:10.1002/9781444315783.ch38

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, UK: Psychology Press. doi:10.4324/9780203841228

De Groot, A. M. B. (2017). *Nederlands moet. Over meertaligheid en de verengelsing van het universitaire onderwijs*. Valedictory lecture, University of Amsterdam.

De Groot, A., Jurgens, E., Rawie, J. P., & Verbrugge, A. (2018). *Verengelsing is geen geneuzel, minister Van Engelshoven*. Retrieved from https://www.volkskrant.nl/columns-opinie/verengelsing-is-geen-geneuzel-minister-van-engelshoven~b56a370e/

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 Koning, B. B., Loyens, S. M. M., Rikers, R. M. J. P., Smeets, G., & Van der Molen, H. T. (2012). Generation Psy: Student characteristics and academic achievement in a three-year problem-based learning bachelor program. *Learning and Individual Differences, 22*(3), 313–323. doi:10.1016/j.lindif.2012.01.003

De la Fuente, M. J. (2002). Negotiation and oral acquisition of L2 vocabulary. The roles of input and output in the receptive and productive acquisition of words. *Studies in Second Language Acquisition, 24*(1), 81–112. doi:10.1017/S0272263102001043

De Wachter, L., Heeren, J., Marx, S., & Huyghe, S. (2013). Taal: Noodzakelijke, maar niet enige voorwaarde tot studiesucces. Correlatie tussen resultaten van een taalvaardigheidstoets en slaagcijfers bij eerstejaarsstudenten aan de KU Leuven. *Levende Talen Tijdschrift, 14*(4), 28–36.

Dewaele, J.-M. (2005). The effect of type of acquisition context on perception and self-reported use of swearwords in L2, L3, L4 and L5. In A. Housen & M. Pierrard (Eds.), *Investigations in instructed second language acquisition* (pp. 531–559). Berlin: De Gruyter. doi:10.1515/9783110197372.4.531

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science, 11*(3), 189–228.

Donkaewbua, S. (2009). *Vocabulary learning through listening in another language.* Saarbrücken, Germany: Lambert Academic Publishing.

Doughty, C. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 206–257). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139524780.010

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, UK: Cambridge University Press.

Douglas, S. R. (2010). *Non-native English speaking students at university: Lexical richness and academic success* (Doctoral dissertation). Retrieved from https://dspace.ucalgary.ca/handle/1880/48195

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*, 52–64. doi:10.2307/2282330

Duquette, L. (1993). *L'etude de l'apprentissage du vocabulaire en contexte par l'ecoute d'un dialogue scenarise en francais langue seconde* [The study of vocabulary learning in context by listening to a dialogue in scenario form in French as a second language] (Report No. CIRAL-B-187). Quebec, Canada: International Center for Research on Language Planning.

Ebbinghaus, H. (1885, translated 1913, reprinted 2011). *Memory: A contribution to experimental psychology*. Eastford, CT: Martino Fine Books. doi:10.1037/10011-000

Eckerth, J., & Tavakoli, P. (2012). The effects of word exposure frequency and elaboration of word processing on incidental L2 vocabulary acquisition through reading. *Language Teaching Research, 16*(2), 227–252. doi:10.1177/1362168811431377

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

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

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, R., & He, X. (1999). The roles of modified input and output in the incidental acquisition of word meanings. *Studies in Second Language Acquisition, 21*(2), 285–301. doi:10.1017/s0272263199002077

Ellis, R., & Heimbach, R. (1997). Bugs and birds: Children's acquisition of second language vocabulary through interaction. *System, 25*(2), 247–259. doi:10.1016/s0346-251x(97)00012-2

Ellis, R., Tanaka, Y., & Yamazaki, A. (1994). Classroom interaction, comprehension, and the acquisition of L2 word meanings. *Language Learning, 44*(3), 449–491. doi:10.1111/j.1467-1770.1994.tb01114.x

Eysenck, M. W. (1982): Incidental learning and orienting tasks. In C. R. Puff (Ed.), *Handbook of research methods in human memory and cognition* (pp. 197–228). New York, NY: Academic Press.

Fakeye, D. O., & Ogunsiji, Y. (2009). English language proficiency as a predictor of academic achievement among EFL students in Nigeria. *European Journal of Scientific Research, 37*(3), 490–495.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Field, A. (2009). *Discovering Statistics Using SPSS* (3rd ed.). London, UK: SAGE.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. London, UK: SAGE.

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy, 98*, 19–38. doi:10.1016/j.brat.2017.05.013

Fonteyne, L., De Fruyt, F., & Duyck, W. (2014). To fail or not to fail? Identifying students at risk by predicting academic success. In M. F. Freda (Ed.), *Reflexivity in higher education: Research and models of intervention for underachieving students* (pp. 345–356). Rome, Italy: Aracne. doi:10.4399/978885487014727

Frear, M. W., & Bitchener, J. (2015). The effects of cognitive task complexity on writing complexity. *Journal of Second Language Writing, 30*, 45–57. doi:10.1016/j.jslw.2015.08.009

Gass, S. (1999). Discussion: Incidental vocabulary learning. *Studies in Second Language Acquisition, 21*(2), 319–333.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum. doi:10.4324/9781410606006

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Gelman, A., & Su, Y.-S. (2018). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models [Computer software]. Retrieved from https://cran.r-project.org/web/packages/arm/index.html

Gelman, A. et al. (2016). Package 'arm' [Computer software]. Retrieved from https://cran.r-project.org/web/packages/arm/index.html

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

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye tracking. *Studies in Second Language Acquisition, 35*(3), 483–517. doi:10.1017/S0272263113000119

Godfroid, A., Housen, A., & Boers, F. (2010). A procedure for testing the Noticing Hypothesis in the context of vocabulary acquisition. In M. Pütz & L. Sicola (Eds.), *Inside the learner's mind: Cognitive processing and second language acquisition* (pp. 169–197). Amsterdam: John Benjamins. doi:10.1075/celcr.13.14god

Goldschneider, J. M., & DeKeyser, R. M. (2001). Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning, 51*(1), 1–50. doi:10.1111/1467-9922.00147

Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research, 29*(3), 311–343. doi:10.1177/0267658312461497

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

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition, 40*(4), 505–513. doi:10.3758/s13421-011-0174-0

Gruber, M. J., Gelman, B. D., & Ranganath, C. (2014). States of curiosity modulate hippocampus-dependent learning via the dopaminergic circuit. *Neuron, 84*(2), 486–496. doi:10.1016/j.neuron.2014.08.060

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.

Hanaoka, O. (2007). Output, noticing, and learning: An investigation into the role of spontaneous attention to form in a four-stage writing task. *Language Teaching Research, 11*(4), 459–479. doi:10.1177/1362168807080963

Hanaoka, O., & Izumi, S. (2012). Noticing and uptake: Addressing pre-articulated covert problems in L2 writing. *Journal of Second Language Writing, 21*(4), 332–347. doi:10.1016/j.jslw.2012.09.008

Harsch, C., & Hartig, J. (2016). Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills. *Language Testing, 33*(4), 555–575. doi:10.1177/0265532215594642

Hatami, S. (2017). The differential impact of reading and listening on L2 incidental acquisition of different dimensions of word knowledge. *Reading in a Foreign Language, 29*(1), 61–85.

Hedges, L. V, Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39–65. doi:10.1002/jrsm.5

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Cambridge, MA: Academic Press. doi:10.1016/b978-0-08-057065-5.50001-4

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance* (Doctoral dissertation). Retrieved from http://hdl.handle.net/11370/6ff6bbca-842f-4a90-9c6e-a0d3cce748da

Hellekjaer, G. O. (2010). Assessing lecture comprehension in Norwegian English-medium higher education. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *Language use and language learning in CLIL classrooms* (pp. 233–258). Amsterdam: John Benjamins. doi:10.1075/aals.7

Hermans, F. (2017). *Studenten willen geen les meer in steenkolen-Engels van docenten*. Retrieved from https://www.gelderlander.nl/nijmegen-e-o/studenten-willen-geen-les-meer-in-steenkolen-engels-van-docenten~a5a997b1/

Högel, J., Schmid, W., & Gaus, W. (1994). Robustness of the standard deviation and other measures of dispersion. *Biometrical Journal, 36*(4), 411–427. doi:10.1002/bimj.4710360403

Hopkins, W. G. (2002). *A new view of statistics*. Retrieved from http://www.sportsci.org/resource/stats/effectmag.html

Horstmann, G., & Herwig, A. (2016). *Novelty biases attention and gaze in a surprise trial. Attention, Perception, & Psychophysics, 78*, 69–77. doi:10.3758/s13414-015-0995-1

Howard, M. (2005). Second language acquisition in a study abroad context: A comparative investigation of the effects of study abroad and foreign language instruction on the L2 learner's grammatical development. In A. Housen & M. Pierrard (Eds.), *Investigations in instructed second language acquisition* (pp. 495–530). Berlin: De Gruyter. doi:10.1515/9783110197372.4.495

Hsu, C.-K., Hwang, G.-J., Chang, Y.-T., & Chang, C.-K. (2013). Effects of video caption modes on English listening comprehension and vocabulary acquisition using handheld devices. *Educational Technology & Society, 16*(1), 403–414.

Huang, S., Willson, V., & Eslami, Z. (2012). The effects of task involvement load on L2 incidental vocabulary learning: A meta-analytic study. *The Modern Language Journal, 96*(4), 544–557. doi:10.1111/j.1540-4781.2012.01394.x

Huberts, D. (n.d.). *Update: Incoming student mobility in Dutch higher education 2016-17*. Retrieved from https://www.nuffic.nl/documents/393/update-incoming-student-mobility-in-dutch-higher-education-2016-17.pdf

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, UK: Blackwell. doi:10.1002/9780470756492.ch12

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

Huygen, F. (2017). *Opinie: Engels als voertaal vernielt het hoger onderwijs*. Retrieved from https://www.volkskrant.nl/columns-opinie/opinie-engels-als-voertaal-vernielt-het-hoger-onderwijs~b7fd8359/

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review, 105*(4), 765–789. doi:10.1017/s0003055411000414

Ishak, K. J., Platt, R. W., Joseph, L., & Hanley, J. A. (2008). Impact of approximating or ignoring within-study covariances in multivariate meta-analyses. *Statistics in Medicine, 27*(5), 670–686. doi:10.1002/sim.2913

Izumi, S. (2013). Noticing and L2 development: Theoretical, empirical, and pedagogical issues. In J. M. Bergsleithner, S. Nagem Frota, & J. K. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 37–50). Honolulu, HI: National Foreign Language Resource Center.

Izumi, S., & Bigelow, M. (2000). Does output promote noticing and second language acquisition? *TESOL Quarterly, 34*(2), 239–278. doi:10.2307/3587952

Izumi, S., Bigelow, M., Fujiwara, M., & Fearnow, S. (1999). Testing the Output Hypothesis: Effects of output on noticing and second language acquisition. *Studies in Second Language Acquisition, 21*(3), 421–452. doi:10.1017/s0272263199003034

Izumi, Y., & Izumi, S. (2004). Investigating the effects of oral output on the learning or relative clauses in English: Issues in the psycholinguistic requirements for effective output tasks. *The Canadian Modern Language Review, 60*(5), 587–609. doi:10.3138/cmlr.60.5.587

Jalbert, A., Neath, I., Bireta, T. J., & Surprenant, A. M. (2011). When does length cause the word length effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(2), 338–353. doi:10.3758/s13421-011-0094-z

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning, 64*(1), 160–212. doi:10.1111/lang.12034

Joe, Y., & Lee, H.-K. (2013). Does English-medium instruction benefit students in EFL contexts? A case study of medical students in Korea. *The Asia-Pacific Education Researcher, 22*(2), 201–207. doi:10.1007/s40299-012-0003-7

Johnston, W. A., Hawley, K. J., Plewe, S. H., Elliott, J. M., & Dewitt, M. J. (1990). Attention capture by novel stimuli. *Journal of Experimental Psychology: General, 119*(4), 397–411. doi:10.21236/ada221394

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T., & Camerer, C. F. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science, 20*(8), 963–974. doi:10.2139/ssrn.1308286

Karakaş, A., & Sariçoban, A. (2012). The impact of watching subtitled animated cartoons on incidental vocabulary learning of ELT students. *Teaching English with Technology, 12*(4), 3–15.

Keck, C., Iberri-Shea, G., Tracy-Ventura, N., & Wa-Mbaleka, S. (2006). Investigating the empirical link between task-based interaction and acquisition. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 91–131). Amsterdam: John Benjamins. doi:10.1075/lllt.13

Kleinjan, G. -J. (2017). *Steenkolenengels bij docenten: 'Experts die details en nuances missen. Zo jammer'.* Retrieved from https://www.trouw.nl/home/steenkolenengels-bij-docenten-experts-die-details-en-nuances-missen-zo-jammer-~ade0943b/

KNAW (2017). *Nederlands en/of Engels? Taalkeuze met beleid in het Nederlands hoger onderwijs. Amsterdam: KNAW.* Retrieved from https://www.knaw.nl/en/news/publications/nederlands-en-of-engels

Knoch, U., Rouhshad, A., & Storch, N. (2014). Does the writing of undergraduate ESL students develop after one year of study in an English-medium university? *Assessing Writing, 21*, 1–17. doi:10.1016/j.asw.2014.01.001

Knoch, U., Rouhshad, A., Oon, S. P., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university? *Journal of Second Language Writing, 28*, 39–52. doi:10.1016/j.jslw.2015.02.005

Koch, E., De Vos, J. F., Lemhöfer, K., Housen, A., & Godfroid, A. (2019). *Learning second language morphosyntax in dialogue under explicit and implicit conditions: An experimental study.* Manuscript in preparation.

Koenig, A. M., Eagly, A. H., Mitchell, A. A., & Ristikari, T. (2011). Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological Bulletin, 137*(4), 616–642. doi:10.1037/e617292010-001

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods, 2*(1), 61–76. doi:10.1002/jrsm.35

Koolstra, C. M., & Beentjes, J. W. (1999). Children's vocabulary acquisition in a foreign language through watching subtitled television programs at home. *Educational Technology Research and Development, 47*(1), 51–60. doi:10.1007/BF02299476

Kooreman, A., & Luteijn, F. (1987). *Groninger intelligentie test: Schriftelijke verkorte vorm (GITvv).* Lisse: Swets & Zeitlinger.

Kornell, N., Jensen Hays, M., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(4), 989–998. doi:10.1037/a0015729

Kwon, S. H. (2006). *Roles of output and task design on second language vocabulary acquisition* (Doctoral dissertation). Retrieved from http://etd.fcla.edu/UF/UFE0014501/kwon_s.pdf

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4*, article 863. doi:10.3389/fpsyg.2013.00863

Lakens, D. (2016). *Why you don't need to adjust your α level for all tests you'll do in your lifetime* [Blog post]. Retrieved from http://daniellakens.blogspot.com/2016/02/why-you-dont-need-to-adjust-you-alpha.html

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

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *The Canadian Modern Language Review, 59*(4), 567–588. doi:10.3138/cmlr.59.4.567

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22*(1), 1–26. doi:10.1093/applin/22.1.1

Lei, J., & Hu, G. (2014). Is English-medium instruction effective in improving Chinese undergraduate students' English competence? *International Review of Applied Linguistics in Language Teaching, 52*(2), 99–126. doi:10.1515/iral-2014-0005

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

Lemmouh, Z. (2008). The relationship between grades and the lexical richness of student essays. *Nordic Journal of English Studies, 7*(3), 163–180.

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software, 69*(1), 1–33. doi:10.18637/jss.v069.i01

Leow, R. P. (1999). The role of attention in second/foreign language classroom research: Methodological issues. *In Papers from the 2nd Hispanic Linguistics Symposium*, 60–71.

Leow, R. P. (2000). A study of the role of awareness in foreign language behavior. Aware versus unaware learners. *Studies in Second Language Acquisition, 22*(4), 557–584. doi:10.1017/s0272263100004046

Leow, R. P. (2015). *Explicit learning in the L2 classroom: A student-centered approach*. New York, NY: Routledge. doi:10.4324/9781315887074

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

Li, J., & Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity, 95*(3), 221–227. doi:10.1038/sj.hdy.6800717

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning, 60*(2), 309–365. doi:10.1111/j.1467-9922.2010.00561.x

Litman, J. A. (2008). Interest and deprivation factors of epistemic curiosity. *Personality and Individual Differences, 44*(7), 1585–1595. doi:10.1016/j.paid.2008.01.014

Lizzini, O., Martijn, M., Munk, R., & De Regt, H. (2017). *Academisch onderwijs kan niet zonder het Engels*. Retrieved from https://www.volkskrant.nl/columns-opinie/academisch-onderwijs-kan-niet-zonder-het-engels~bcc9ca31/

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

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2), 190–208. doi:10.1111/j.1540-4781.2011.01232_1.x

Lu, X. (2013). Lexical Complexity Analyzer [Computer software]. Retrieved from http://www.personal.psu.edu/xxl13/downloads/lca.html

Macaro, E., Curle, S., Pun, J., An, J., & Dearden, J. (2018). A systematic review of English medium instruction in higher education. *Language Teaching, 51*(1), 36–76. doi:10.1017/S0261444817000350

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 407–452). Oxford, UK: Oxford University Press.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58*, 593–614. doi:10.1146/annurev.psych.58.110405.085542

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

Maex, K. (2017). *Wij kiezen voor Engels- én Nederlandstalig onderwijs*. Retrieved from https://www.volkskrant.nl/columns-opinie/wij-kiezen-voor-engels-en-nederlandstalig-onderwijs-~b5fa2e95/

Mahmoudabadi, Z., Soleimani, H., Jafarigohar, M., & Iravani, H. (2015). The effect of sequence of output tasks on noticing vocabulary items and developing vocabulary knowledge of Iranian EFL learners. *International Journal of Asian Social Science, 5*, 18–30. doi:10.18488/journal.1/2015.5.1/1.1.18.30

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

McDonald, B. (2002). A teaching note on Cook's distance – A guideline. *Research Letters in the Information and Mathematical Sciences, 3*, 127–128.

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

Meara, P., & Ingle, S. (1986). The formal representation of words in an L2 speaker's lexicon. *Second Language Research, 2*(2), 160–171. doi:10.1177/026765838600200203

Medina, S. L. (1990). The effects of music upon second language vocabulary acquisition. *The Annual Meeting of the Teachers of English to Speakers of Other Languages*. Retrieved from ERIC database. (ED 352-834)

Mickan, A., McQueen, J. M., & Lemhöfer, K. (2019). *The role of between-language competition in foreign language attrition*. Manuscript submitted for publication.

Mickan, A., McQueen, J. M., Piai, V., & Lemhöfer, K. (2018, August). *Neural correlates of between-language competition in foreign language attrition*. Poster presented at the 10th meeting of the Society for the Neurobiology of Language, Québec City, Canada.

Mitchell, R., & Myles, F. (2004). *Second language learning theories* (2nd ed.). London, UK: Hodder Arnold. doi:10.4324/9780203770658

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.

Montero Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014). Effects of captioning on video comprehension and incidental vocabulary learning. *Language Learning & Technology, 18*(1), 118–141.

Montero Perez, M., Peters, E., & Desmet, P. (2015). Enhancing vocabulary learning through captioned video: An eye-tracking study. *The Modern Language Journal, 99*(2), 308–328. doi:10.1111/modl.12215

Montero Perez, M., Peters, E., & Desmet, P. (2018). Vocabulary learning through viewing video: The effect of two enhancement techniques. *Computer Assisted Language Learning, 31*(1-2), 1–26. doi:10.1080/09588221.2017.1375960

Montero Perez, M., Van Den Noortgate, W., & Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System, 41*(3), 720–739. doi:101016/j.system.2013.07.013

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*(2), 364–386. doi:10.1177/1094428106291059

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*(1), 105–125. doi:10.1037//1082-989X.7.1.105

Nagata, H., Aline, D., & Ellis, R. (1999). Modified input, language aptitude and the acquisition of word meanings. In R. Ellis (Ed.), *Learning a second language through interaction* (pp. 133–149). Amsterdam: John Benjamins. doi:10.1075/sibil.17.09nag

Nakata, T. (2016). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition, 39*(4), 653–679. doi:10.1017/S0272263116000280

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139524759

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle.

Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning, 65*(2), 470–476. doi:10.1111/lang.12104

Norris, J. M., Ross, S. J., & Schoonen, R. (2015). Improving second language quantitative research. *Language Learning, 65*(S1), 1–8. doi:10.1111/lang.12110

Nyholt, D. R. (2004). A simple correction for multiple testing for SNPs in linkage disequilibrium with each other. *The American Journal of Human Genetics, 74*(4), 765–769. doi:10.1086/383251

Nyholt, D. R. (2015). *Matrix Spectral Decomposition (matSpD) - Estimate the equivalent number of independent variables in a correlation (r) matrix*. Retrieved from https://neurogenetics.qimrberghofer.edu.au/matSpD/

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

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics, 30*, 85–110. doi:10.1017/S0267190510000115

Paradis, M. (2004). *A neurolinguistic theory of bilingualism*. Amsterdam: John Benjamins. doi:10.1075/sibil.18

Park, E. S. (2007). *Learner-generated noticing of L2 input: An exploratory study* (Unpublished doctoral dissertation), Teachers College, Columbia University, US.

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

Peters, E. (2007a). Manipulating L2 learners' online dictionary use and its effect on L2 word retention. *Language Learning & Technology, 11*(2), 45–67.

Peters, E. (2007b). The influence of task instruction on vocabulary acquisition and reading comprehension. In M. P. Garcia Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 178–198). Clevedon, UK: Multilingual Matters.

Peters, E. (2014). The effects of repetition and time of post-test administration on EFL learners' form recall of single words and collocations. *Language Teaching Research, 18*(1), 75–94. doi:10.1177/1362168813505384

Peters, E., Hulstijn, J. H., Sercu, L., & Lutjeharms, M. (2009). Learning L2 German vocabulary through reading: The effect of three enhancement techniques compared. *Language Learning, 59*(1), 113–151. doi:10.1111/j.1467-9922.2009.00502.x

Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition, 40*(3), 551–557. doi:10.1017/S0272263117000407

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

Plonsky, L. (2013). Study quality in SLA. An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition, 35*(4), 655–687. doi:10.1017/S0272263113000399

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. doi:10.1111/lang.12079

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13*, 25–42. doi:10.1146/annurev.ne.13.030190.000325

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

Python Software Foundation (2018). Python, version 3.6 [Computer software]. Retrieved from http://www.python.org

Quené, H., & Van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication, 43*(1-2), 103–121. doi:10.1016/j.specom.2004.02.004

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

Ramachandra, V., Rickenbach, B., Ruda, M., LeCureux, B., & Pope, M. (2010). Fast mapping in healthy young adults: The influence of metamemory. *Journal of Psycholinguistic Research, 39*(3), 213–224. doi:10.1007/s10936-009-9133-3

Read, J. (2000). Assessing vocabulary. Cambridge, UK: Cambridge University Press. doi:10.1017/cbo9780511732942

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

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied, 15*(3), 243–257. doi:10.1037/a0016496

Rienties, B. C., Beausaert, S. A. J., Grohnert, T., Niemantsverdriet, S., & Kommers, P. (2012). Understanding academic performance of international students: The role of ethnicity, academic and social integration. *Higher Education, 63*(6), 685–700. doi:10.1007/s10734-011-9468-1

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

Rodgers, M. P. H. (2013). *English language learning through viewing television: An investigation of comprehension, incidental vocabulary acquisition, lexical coverage, attitudes, and captions* (Doctoral dissertation). Retrieved from http://hdl.handle.net/10063/2870

Rogier, D. (2012). *The effects of English-medium instruction on language proficiency of students enrolled in higher education in the UAE* (Doctoral dissertation). Retrieved from http://hdl.handle.net/10036/4482

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

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

Sampling variance for meta-analysis one-sample data (2016) [Online forum comment]. Retrieved from http://stats.stackexchange.com/questions/226836/sampling-variance-for-meta-analysis-one-sample-data/

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139524780.003

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158. doi:10.1093/applin/11.2.129

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

Sercu, L., De Wachter, L., Peters, E., Kuiken, F., & Vedder, I. (2006). The effect of task complexity and task conditions on foreign language development and performance: Three empirical studies. ITL: International *Journal of Applied Linguistics, 152*, 55–84. doi:10.2143/ITL.152.0.2017863

Serrano, R., Tragant, E., & Llanes, À. (2012). A longitudinal analysis of the effects of one year abroad. *Canadian Modern Language Review, 68*(2), 138–163. doi:10.3138/cmlr.68.2.138

Shetter, W. Z. (1959). The Dutch diminutive. *The Journal of English and Germanic Philology, 58*, 75–90.

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

Singleton, D., & Ryan, L. (2004). *Language acquisition: The age factor* (2nd ed.). Clevedon, UK: Multilingual Matters. doi:10.21832/9781853597596

Snow, C. E., & Hoefnagel-Höhle, M. (1978). The critical period for language acquisition: Evidence from second language learning. *Child Development, 49*(4), 1114–1128. doi:10.2307/1128751

Song, M.-J., & Suh, B.-R. (2008). The effects of output task types on noticing and learning of the English past counterfactual conditional. *System, 36*(2), 295–312. doi:10.1016/j.system.2007.09.006

Statistics How To (2017). *Pooled standard deviation*. Retrieved from http://www.statisticshowto.com/pooled-standard-deviation/

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Rowley, MA: Newbury House.

Swain, M. (1993). The Output Hypothesis: Just speaking and writing aren't enough. *The Canadian Modern Language Review, 50*(1), 158–164.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principles and practice in applied linguistics: Studies in honour of H.G. Widdowson* (pp. 125–144). Oxford, UK: Oxford University Press.

Swain, M. (1998). Focus on form through conscious reflection. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 64–81). Cambridge, UK: Cambridge University Press.

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

Sydorenko, T. (2010). Modality of input and vocabulary acquisition. *Language Learning & Technology, 14*(2), 50–73.

Tanaka, K., & Ellis, R. (2003). Study-abroad, language proficiency, and learner beliefs about language learning. *JALT Journal, 25*(1), 63–85.

Tatzl, D., & Messnarz, B. (2013). Testing foreign language impact on engineering students' scientific problem-solving performance. *European Journal of Engineering Education, 38*(6), 620–630. doi:10.1080/03043797.2012.719001

Teuling, I. (2017). *Zorgen over kwaliteit Engelstalig onderwijs universiteiten en hogescholen*. Retrieved from https://www.volkskrant.nl/nieuws-achtergrond/zorgen-over-kwaliteit-engelstalig-onderwijs-universiteiten-en-hogescholen~bcce8e6c/

The Effect Size (n.d.). Retrieved from psych.unl.edu/psycrs/971/meta/effect_sizes.ppt

Tomlin, R. S., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition, 16*(2), 183–203. doi:10.1017/s0272263100012870

Toya, M. (1993). *Form of explanation in modification of listening input in L2 vocabulary learning*. Occasional Papers Series, Department of English as a Second Language, University of Hawai'i, United States.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics, 5*(2), 99–114. doi:10.2307/3001913

Uggen, M. S. (2012). Reinvestigating the noticing function of output. *Language Learning, 62*(2), 506–540. doi:10.1111/j.1467-9922.2012.00693.x

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 den Bosch, A., Busser, G. J., Daelemans, W., & Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In F. van Eynde, P. Dirix, I. Schuurman, & V. Vandeghinste (Eds.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting* (pp. 99–114).

Van den Broek, G. S. E., Takashima, A., Segers, E., & Verhoeven, L. (2018). Contextual richness and word learning: Context enhances comprehension but retrieval enhances retention. *Language Learning, 68*(2), 546–585. doi:10.1111/lang.12285.

Van der Westen, W., & Wijsbroek, D. (2011). Slecht in taal, slecht in studie? Resultaten van een onderzoek naar de relatie tussen taalvaardigheid en studiesucces. In A. Mottart (Chair), *Vijfentwintigste Conferentie Het Schoolvak Nederlands* (pp. 118–123).

Van Houtven, T., Peters, E., & El Morabit, Z. (2010). Hoe staat het met de taal van studenten? Exploratieve studie naar begrijpend lezen en samenvatten bij instromende studenten in het Vlaamse hoger onderwijs. *Levende Talen Tijdschrift, 11*(3), 29–44.

Van Oostendorp, M. (2017). *Voer de verengelsing nog maar wat verder*. Retrieved from https://www.neerlandistiek.nl/2017/07/voer-de-verengelsing-nog-maar-wat-door/

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

Vander Beken, H., Woumans, E., & Brysbaert, M. (2018). Studying texts in a second language: No disadvantage in long-term recognition memory. *Bilingualism: Language and Cognition, 21*(4), 826–838. doi:10.1017/S1366728917000360

Vasterman, P. (2017). *Keuze voor Engels schaadt de inhoud op de universiteit*. Retrieved from https://www.nrc.nl/nieuws/2017/11/13/keuze-voor-engels-schaadt-de-inhoud-op-de-universiteit-13989528-a1580911

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

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. doi:10.18637/jss.v036.i03

Viechtbauer, W. (2017). *Konstantopoulos (2011)*. Retrieved from: http://www.metafor-project.org/doku.php/analyses:konstantopoulos2011

Vinke, A. A. (1995). *English as the medium of instruction in Dutch engineering education* (Doctoral dissertation). Retrieved from Research Repository TU Delft (Accession no. uuid:491b55f9-fbf9-4650-a44d-acb9af8412a8)

Vinke, A. A., Snippe, J., & Jochems, W. (1998). English-medium content courses in non-English higher education: A study of lecturer experiences and teaching behaviours. *Teaching in Higher Education, 3*(3), 383–394. doi:10.1080/1356215980030307

Vulchanova, M., Aurstad, L. M., Kvitnes, I. E., & Eshuis, H. (2015). As naturalistic as it gets: Subtitles in the English classroom in Norway. *Frontiers in Psychology, 5*, article 1510. doi:10.3389/fpsyg.2014.01510

Wächter, B., & Maiworm, F. (2014). *English-taught programmes in European higher education: The state of play in 2014*. Bonn: Lemmens.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition, 27*(1), 33–52. doi:10.1017/S0272263105050023

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*(1), 46–65. doi:10.1093/applin/aml048

Wilcox, R. R. (2005). *Robust estimation and hypothesis testing*. Burlington, MA: Elsevier Academic Press.

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

Williams, J. (1999). Learner-generated attention to form. *Language Learning, 49*(4), 583–625. doi:10.1111/0023-8333.00103.

Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology, 14*(1), 65–86.

Winter, B. (2013). *Linear models and linear mixed effects models in R with linguistic applications* [arXiv:1308.5499]. Retrieved from http://arxiv.org/pdf/1308.5499.pdf

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity* [Report No. 17]. Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Wright, D. B., London, K., & Field, A. P. (2011). Using bootstrap estimation and the plug-in principle for clinical psychology data. *Journal of Experimental Psychopathology, 2*(2), 252–270. doi:10.5127/jep.013611

Yang, W. (2015). Content and language integrated learning next in Asia: Evidence of learners' achievement in CLIL education from a Taiwan tertiary degree programme. International *Journal of Bilingual Education and Bilingualism, 18*(4), 361–382. doi:10.1080/13670050.2014.904840

Yang, W., Lu, X., & Cushing Weigle, S. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing, 28*, 53–67. doi:10.1016/j.jslw.2015.02.002

Yeung, S. S., Ng, M., & King, R. B. (2016). English vocabulary instruction through storybook reading for Chinese EFL kindergarteners: Comparing rich, embedded, and incidental approaches. *Asian EFL Journal, 18*(2), 89–112.

Yuksel, D., & Tanriverdi, B. (2009). Effects of watching captioned movie clip on vocabulary development of EFL learners. *Turkish Online Journal of Educational Technology, 8*(2), 48–54.

Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis. *Computer Assisted Language Learning, 24*(1), 39–58. doi:10.1080/09588221.2010.523285

Zijlmans, L., Neijt, A., & Van Hout, R. (2016). The role of second language in higher education: A case study of German students at a Dutch university. *Language Learning in Higher Education, 6*(2), 473–493. doi:10.1515/cercles-2016-0026

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

NEDERLANDSE SAMENVATTING

Een tweede taal leren

In Nederland krijgt iedereen op school les in vreemde talen: in elk geval Engels, en meestal ook Duits, Frans, of nog andere talen. In de taalwetenschap gebruiken we vaak de term *tweede taal* om te verwijzen naar alle vreemde talen die iemand naast zijn of haar moedertaal probeert te leren, ook al gaat het letterlijk gezien misschien wel om een derde of vierde taal. Tweede talen leer je niet alleen op school. De meeste Nederlanders komen ook buiten school regelmatig in aanraking met vreemde talen, bijvoorbeeld op vakantie (denk aan het kopen van stokbrood in een Franse bakkerij). Daarnaast kijken veel mensen wel eens naar een Engelstalige film of serie. Ook in zulke situaties kunnen mensen hun vaardigheid in de tweede taal verbeteren. Leren in dit soort situaties noemen we *naturalistisch* leren, en de lerende persoon wordt een *leerder* genoemd.

In vergelijking met taalonderwijs zoals dat op school of bij een taalcursus plaatsvindt, verloopt naturalistisch leren over het algemeen heel anders. Naturalistisch taalleren vindt vaak plaats in een informele context, en zonder expliciete taalinstructie (een voorbeeld van een expliciete instructie is: “een vrouwelijk paard heet een *merrie*”). Er is relatief weinig bekend over hoe het leerproces van tweede-taalleerders in naturalistische situaties verloopt, en hoeveel ze in dit soort situaties kunnen leren. Dat komt omdat dit soort leerprocessen vaak lastig te onderzoeken zijn: Onderzoekers kunnen geen controle uitoefenen over de leersituatie (bijvoorbeeld: welke woorden krijgen leerders te horen, en hoe vaak), en hebben vaak niet eens toegang tot deze informatie. Zulke controle heb je als onderzoeker wel als je een strikt gecontroleerd laboratoriumonderzoek opzet, maar dat is dan weer minder naturalistisch. Dit is een dilemma binnen het onderzoek naar naturalistische taalverwerving.

Waar gaat dit proefschrift over?

Het doel van dit proefschrift was om meer te weten te komen over de naturalistische verwerving van woordenschat in een tweede taal, en daarbij zowel aandacht te besteden aan goede experimentele controle alsook aan de natuurlijkheid van de leersituatie. Ik gebruik de term *input* voor alle taal waaraan een leerder wordt blootgesteld, bijvoorbeeld geschreven taal in een boek, of gesproken taal in een film of gesprek. In Hoofdstukken 2 tot en met 4 van dit proefschrift heb ik me specifiek gericht op leren op basis van gesproken input, omdat de input in naturalistische leersituaties vaak gesproken is, en hier bovendien minder over bekend is dan over leren van geschreven input. Hoofdstuk 2 is een overzichtsstudie van eerdere onderzoeken over het leren van woorden uit gesproken input, terwijl ik in Hoofdstukken 3 en 4 experimenten beschrijf die ik zelf heb uitgevoerd op dit gebied.

Hoofdstukken 5 en 6 richten zich op het gebruik van Engels als onderwijstaal aan Nederlandse universiteiten, in dit geval de Radboud Universiteit in Nijmegen. De verwerving van Engels door studenten in deze context is ook naturalistisch: Engels is hier een middel, maar geen doel op zich. In Hoofdstuk 5 heb ik onderzocht hoe de Nederlandse en Engelse taalverwerving van studenten eruitziet gedurende hun eerste studiejaar aan de universiteit.

Hoofdstuk 6 gaat in op de actuele vraag of de studietaal (Nederlands of Engels) invloed heeft op de resultaten die studenten behalen. Nu volgt een uitgebreider overzicht van alle hoofdstukken.

Meta-analyse

Om te beginnen wilde ik op een rijtje zetten wat er uit eerder onderzoek al bekend was over naturalistische woordenschatverwerving in een tweede taal, op basis van gesproken input. Deze overzichtsstudie van eerder onderzoek staat in **Hoofdstuk 2**. Mijn doel was om de resultaten uit eerder onderzoek naar naturalistische woordenschatverwerving in een tweede taal samen te voegen in één grote analyse, genaamd een *meta-analyse*. Ik kwam er echter al vrij snel achter dat dit niet goed mogelijk was, omdat het relatief schaarse onderzoek naar naturalistische woordenschatverwerving te divers was, en uitgevoerd met te weinig experimentele controle. Daarom heb ik me gericht op *incidentele* woordenschatverwerving, dat een aantal belangrijke kenmerken deelt met naturalistische woordenschatverwerving en daarom ook bij kan dragen aan onze kennis over dit onderwerp.

De kenmerken die beide manieren van leren (incidenteel en naturalistisch) met elkaar delen is dat er leren plaatsvindt zonder dat dit het hoofddoel is van de activiteit waarmee de leerder bezig is. Zo kun je bijvoorbeeld nieuwe woorden leren terwijl je naar een film kijkt; de woordenschatverwerving is dan in principe niet het hoofddoel van het filmkijken. Een tweede gedeeld kenmerk is dat leerders niet verwachten dat ze overhoord zullen worden, en zich dus ook niet op een test aan het voorbereiden zijn. Het kenmerk waarin incidenteel en naturalistisch leren wél van elkaar verschillen is dat men ervan uitgaat dat leerders bij incidenteel leren niet de bewuste intentie hebben om iets te leren. In naturalistische leersituaties kan zo'n intentie zowel aan- als afwezig zijn.

Mijn meta-analyse liet zien dat leerders veel woorden uit gesproken input kunnen leren in incidentele leersituaties. Volwassen leerders presteerden beter dan kinderen, wat interessant is omdat kinderen een tweede taal op de lange termijn vaak beter onder de knie krijgen dan volwassenen. Waarschijnlijk heeft het er onder andere mee te maken dat volwassenen beter ontwikkelde leerstrategieën hebben om op de *korte* termijn iets te onthouden. In deze meta-analyse waren de volwassen deelnemers misschien ook gemotiveerder dan de kinderen. Een ander resultaat was dat de deelnemers meer leerden in interactieve dan in niet-interactieve leersituaties, en dat ze het makkelijker vonden om nieuwe woorden te herkennen dan om deze zelf te produceren.

Ten slotte heb ik gevonden dat het voor een juiste inschatting van het leereffect belangrijk is dat de leerders in een studie ook vergeleken worden met een groep deelnemers die niet wordt blootgesteld aan input. Dit is belangrijke kennis voor toekomstige onderzoekers die een studie over woordenschatverwerving willen opzetten.

Twee naturalistische experimenten

Na de meta-analyse heb ik twee gecontroleerde experimenten over naturalistisch woordleren uitgevoerd. De deelnemers waren Duitse studenten in Nijmegen, die Nederlands als tweede taal hadden geleerd. Hoewel ze op het moment van deelname geen taalcursussen meer volgden, was hun Nederlandse taalvaardigheid nog steeds in ontwikkeling. De belangrijkste naturalistische aspecten van de experimenten waren dat de deelnemers niet wisten dat ze meededen aan een studie over woordenschatverwerving, en ook dat ze niet wisten dat ik alleen Duitse studenten had benaderd. Ze dachten zelfs dat het onderzoek helemaal niet over taal ging, maar over het inschatten van de prijs van voorwerpen. Tijdens het onderzoek hoorden de deelnemers mij verschillende voorwerpen benoemen en qua prijs vergelijken met andere voorwerpen. Zo werden ze in de gelegenheid gesteld om nieuwe woorden te leren. Als ze daarna zelf een voorwerp moesten benoemen en met een ander voorwerp vergelijken qua prijs, kon ik vaststellen of ze de naam wel of niet hadden geleerd.

Het eerste van de twee experimenten, beschreven in **Hoofdstuk 3**, was bedoeld om vast te stellen hoeveel nieuwe woorden de deelnemers konden leren in deze naturalistische en tegelijkertijd experimenteel gecontroleerde situatie, en wat de invloed was van meerdere factoren op de leeruitkomsten. Ten eerste heb ik gevonden (zoals verwacht) dat het helpt om woorden vaker te horen. De deelnemers presteerden namelijk beter nadat ze de woorden vier keer hadden gehoord in vergelijking met twee keer. De relatieve vooruitgang was echter het grootst na de eerste twee blootstellingen aan een nieuw woord, in vergelijking met de laatste twee. Verder was het makkelijker voor de deelnemers om nieuwe woorden te leren die verwant waren aan woorden in hun moedertaal. Tussen het moment dat de deelnemers de naam van een voorwerp hoorden, en het moment dat ze dit voorwerp zelf moesten benoemen, kwam steeds ook een aantal andere voorwerpen voorbij. Voor de scores bleek het niet uit te maken hoe groot dit aantal tussenliggende voorwerpen was.

Twintig minuten en zes maanden na het experiment heb ik opnieuw getest hoeveel woorden de deelnemers zich nog konden herinneren. In vergelijking met de tests die al eerder, tijdens de leertaak hadden plaatsgevonden, konden de deelnemers twintig minuten na afloop van het experiment nog ongeveer driekwart van de woorden produceren. Zes maanden na het experiment was dit nog ongeveer een derde. Met name dat laatste resultaat is eigenlijk verrassend goed, gegeven het feit dat er zoveel tijd was verstreken en dat de deelnemers niet wisten dat ze nog een keer getest zouden gaan worden.

Een zeer belangrijke bijdrage van dit eerste experiment was het ontwikkelen van een naturalistische en tegelijkertijd experimenteel gecontroleerde leersituatie, door middel van mijn prijsvergelijkingstaak. Ik heb deze taak in **Hoofdstuk 4** dan ook opnieuw gebruikt om een andere vraag te onderzoeken. Die vraag was of mensen meer nieuwe woorden leren als ze zich bewust zijn van 'gaten' in hun woordenschat. Met andere woorden: Als je je ervan bewust bent dat je een bepaald woord niet kent (wat is bijvoorbeeld het Engelse woord voor *garde*?), zul je dit woord dan beter onthouden als je het op een later moment hoort?

Ik creëerde deze bewustwording door de Duitse deelnemers te vragen om plaatjes van voorwerpen (zoals een garde) in het Nederlands te benoemen. Als dat niet lukt, merkt iemand natuurlijk dat hij of zij voor dit woord een gat in zijn/haar woordenschat heeft. Nadat zo'n bewustwording voor een aantal voorwerpen was gecreëerd, kregen de deelnemers de juiste namen van mij te horen in de prijsvergelijkingstaak. Direct na het experiment testte ik hoeveel van deze namen de deelnemers nog konden produceren. Vijftien minuten later testte ik dit nog een keer. Ik vergeleek deze scores met de scores van deelnemers bij wie *geen* bewustwording van gaten in het woordenschat was gecreëerd. Zij hadden de plaatjes aan het begin van het experiment niet hoeven benoemen. In vergelijking met de wel-bewuste groep had de niet-bewuste groep substantieel minder woorden geleerd.

Een interessante bijkomstigheid was dat er ook binnen de groep van deelnemers die de voorwerpen niet hadden hoeven benoemen in het begin van het experiment, toch een aantal deelnemers zaten die zich wél bewust waren geworden van gaten in hun woordenschatkennis. Ik vergeleek hun leerscores met die van de bewuste deelnemers die de voorwerpen wél hadden benoemd, en vond geen verschil. Dit laat zien dat het niet het (proberen te) benoemen zélf is dat latere woordenschatverwerving faciliteert, maar de bewustwording over welke woorden je nog niet kent in een tweede taal.

Engels als onderwijstaal aan de universiteit

In de periode dat ik aan de bovengenoemde drie studies werkte, was er in de media veel aandacht voor het toenemende gebruik van Engels als onderwijstaal aan Nederlandse universiteiten en hogescholen. Dit debat kent felle voor- en tegenstanders, maar de meeste argumenten die zij gebruiken zijn eerder gebaseerd op meningen dan op wetenschappelijke feiten. Aan de Radboud Universiteit deed zich een goede mogelijkheid voor om te onderzoeken wat nu echt de voor- en nadelen zijn van het gebruik van Engels en Nederlands als onderwijstaal. Vanaf het studiejaar 2016-2017 wordt de bachelor Psychologie namelijk in twee varianten aangeboden: een volledig Engelstalig programma, en een tweetalig Nederlands-Engels programma. In het tweetalige programma worden de colleges en werkgroepen in het Nederlands gegeven, en worden de examens in het Nederlands afgenomen, maar zijn de studiematerialen soms in het Nederlands en soms in het Engels. Door de prestaties van studenten in beide programma's te vergelijken kon ik onderzoeken wat de invloed van de onderwijstaal was op de taalontwikkeling en het studiesucces (bijvoorbeeld de behaalde cijfers) van de studenten.

In **Hoofdstuk 5** heb ik me gericht op de taalontwikkeling, en meer specifiek op de productieve woordenschat van studenten. Dit betrof de diversiteit en complexiteit van de woorden die zij gebruikten in hun geschreven teksten. In dit geval waren die teksten hun antwoorden op open tentamenvragen. Ik heb gebruik gemaakt van drie tentamens: één uit oktober 2016, één uit februari 2017, en één uit april 2017. Aan de hand hiervan kon ik de taalontwikkeling van de studenten door de tijd heen bestuderen. De studenten in het tweetalige programma hadden de tentamenvragen in het Nederlands beantwoord, en de

studenten in het Engelstalige programma in het Engels. Naast de invloed van de studietaal hebben we ook gekeken of er verschillen waren tussen Nederlandse en Duitse studenten.

De resultaten lieten zien dat de Nederlandse studenten niet meer of minder woorden leerden in het Engels dan in het Nederlands, en hetzelfde gold voor de Duitse studenten. Wel was het zo dat de Nederlandse studenten al bij voorbaat een betere productieve woordenschat hadden in het Nederlands dan in het Engels. De productieve woordenschat van Duitse studenten in het Nederlands was even goed als die van de Nederlandse studenten in het Nederlands, wat ik niet had verwacht. Misschien is een aanzienlijk deel van deze Duitse studenten opgegroeid met Nederlands als tweede (moeder)taal, of waren mijn woordenschatmaten niet gevoelig genoeg om subtiele verschillen tussen de twee groepen te kunnen detecteren. Maar de belangrijkste conclusie uit dit hoofdstuk is dat studeren in de ene studietaal (Nederlands) dus niet tot meer of minder taalontwikkeling leidt dan studeren in de andere studietaal (Engels).

In **Hoofdstuk 6** heb ik het 'studiesucces' van de Nederlandse en Duitse studenten in de twee programma's vergeleken. De Nederlandse studenten in het tweetalige programma bleken hogere cijfers te halen dan de Nederlandse studenten in het Engelstalige programma, met een verschil van 0.55 punten op een schaal van 1-10. Dit kan er niet aan liggen dat de Nederlandse studenten in het tweetalige programma slimmer waren dan de Nederlandse studenten in het Engelstalige programma, want op de middelbare school hadden de twee groepen gemiddeld ongeveer dezelfde cijfers behaald. Overigens waren de Nederlandse studenten in het Engelstalige programma op de middelbare school wel al beter in Engels dan de Nederlandse studenten die na hun middelbare schooltijd voor het tweetalige programma zouden kiezen, maar ik heb ervoor gezorgd dat dit verschil geen rol kon spelen in de statistische analyse. De Nederlandse studenten in het tweetalige programma haalden ook hogere cijfers dan de Duitse studenten in beide programma's. Er was geen verschil tussen de vier groepen wat betreft het aantal behaalde studiepunten en hoe vaak de studenten hun studie voortijdig afbraken.

De taalvaardigheid van de studenten (hun productieve woordenschat, zoals gemeten in Hoofdstuk 5) hing niet samen met hun behaalde cijfers. Wel was het zo dat studenten die relatief meer inhoudswoorden gebruikten op hun tentamens (zelfstandig naamwoorden, werkwoorden, bijvoeglijk naamwoorden, en bijwoorden) minder vaak stopten met de studie. Voor dit verband heb ik niet direct een verklaring, maar ik vermoed dat er misschien een onderliggende variabele is, zoals intelligentie of werkgeheugen, die er zowel voor zorgt dat studenten meer inhoudswoorden gebruiken alsook dat ze minder vaak uitvallen. Zo'n verband zou verder onderzocht moeten worden in toekomstig onderzoek.

De conclusie van dit hoofdstuk is dat studenten die in hun moedertaal studeren betere cijfers halen dan studenten die in een tweede taal studeren. Het is nog niet precies duidelijk waar dit aan ligt: Begrijpen zij de colleges of de lesmaterialen beter, en/of kunnen ze hun gedachten beter onder woorden brengen op tentamens? De verklaring zou ook deels bij de docenten kunnen liggen: Nederlands was de moedertaal van de meeste docenten,

en Engels was voor de meeste docenten een tweede taal. Het is dus waarschijnlijk dat de meeste docenten zich beter hebben kunnen uitdrukken in hun Nederlandstalige dan hun Engelstalige colleges. De uiteindelijke verklaring is waarschijnlijk een combinatie van de bovenstaande verklaringen.

Conclusie

In dit proefschrift heb ik naturalistische woordenschatverwerving in een tweede taal vanuit verschillende perspectieven onderzocht. Dit heeft geleid tot nieuwe inhoudelijke inzichten, tot de ontwikkeling van een nieuwe taak om naturalistisch leren in een gecontroleerde labomgeving te onderzoeken, en tot een objectieve analyse van het gebruik van Engels aan Nederlandse universiteiten. Met name het feit dat studenten lagere cijfers halen als ze niet in hun moedertaal studeren geeft aan dat de wenselijkheid van de opmars van het Engels in het hoger onderwijs in twijfel getrokken mag worden.

DANKWOORD

En dan nu het gedeelte in dit proefschrift dat ongetwijfeld het meest gelezen zal worden ;)

Kristin en **Herbert**, ik heb het heel erg gewaardeerd dat er altijd zoveel ruimte was voor mijn eigen ideeën. Op de momenten dat ik wilde afwijken van het oorspronkelijke onderzoeksvoorstel (en dat waren er veel), was dit eigenlijk altijd mogelijk en gaven jullie mij de ruimte om nieuwe statistische technieken uit te proberen en samenwerkingen met andere onderzoekers aan te gaan. **Kristin**, een extra bedankje aan jou voor je grote betrokkenheid bij mijn voortgang en het feit dat je, ondanks je drukke schema, altijd beschikbaar was voor een afspraak als ik iets wilde overleggen.

Van een aantal mensen heb ik cruciale hulp gehad bij de statistische analyses in dit proefschrift. **Louis**, dankjewel dat je de tijd wilde nemen om mijn analyses helemaal uit te pluizen. Ik heb er veel van geleerd! Ook door **Conor** begrijp ik linear mixed-effects models nu veel beter, dankzij onze uitgebreide e-mailconversatie van 96 berichten over en weer, met de toepasselijke titel “nog bezig”. **Pierre**, dankjewel dat je helemaal in het begin van mijn PhD wilde meedenken over de analyses. **Michel**, zonder jouw hulp had ik de meta-analyse niet af kunnen maken (of in elk geval niet in de geruststellende wetenschap dat de statistiek in orde was!).

De twee hoofdstukken over Engelstalig onderwijs zijn tot stand gekomen met de medewerking van het Onderwijsinstituut Psychologie en Kunstmatige Intelligentie, en in het bijzonder met de ondersteuning van **Ruud**, **José**, **Folkert**, **Cristel** en **Eljan**. Docenten **Dennis** en **Jules** leenden een grote hoeveelheid tentamens aan mij uit. **Christiaan**, **Moniek** en **Rick** hebben vele uren besteed aan het overtypen van deze handgeschreven tentamens op de computer, een saaie taak die zij geduldig hebben uitgevoerd. Ten slotte was dit alles niet mogelijk geweest zonder de bereidheid van de **eerstejaarsstudenten Psychologie 2016-2017** om hun data door mij te laten analyseren. Allemaal heel erg bedankt.

In de categorie ‘saaie taken’ bedank ik ook heel graag **Eva**, **Iris** en **Julia**, die samen alle studies in de meta-analyse nog eens hebben doorgelezen om te checken of ik er wel de juiste gegevens uit had gehaald. Dit was heel erg waardevol voor het onderzoek!

Marpessa en ik hebben ruim vier jaar lang lief en leed gedeeld. Ik had me geen beter kantoorgenootje kunnen wensen en ik heb je gemist toen je vanuit Amsterdam begon te werken. **Kasia**, jij was ook een heel fijn kantoorgenootje tijdens mijn eerste jaar en ik vond je aanwezigheid altijd motiverend.

Anne, **Annika**, **Eva**, **Iris**, **Jana**, **Julia**, **Monica**, **Sybrine**, **Willeke** and **Xiaochen** (the members of Kristin's research group), thank you all for listening to me presenting my research for countless times, asking questions and thinking along. It was great to be your colleague both on campus and outside (I'm thinking Mallorca, Ghent, and other places!).

Verschillende mensen hebben me geholpen met kleine dingetjes die veel te veel tijd hadden gekost als ik ze zelf had moeten uitzoeken. **Aaron**, dankjewel voor je eerste hulp bij computerproblemen (en de spontane kopjes thee). **Max**, dankjewel voor dat driehoekje in Photoshop. **Sybrine**, dankjewel dat je me hebt geholpen bij het behalen van mijn BKO en je tips over vacatures.

Fenny, we begonnen als collega's maar al snel gingen onze gesprekken over veel meer dan werk. Ik ben blij dat we zoveel contact hebben gehouden sinds je op reis bent gegaan, en ook dat jij en **Roemer** inmiddels weer terug zijn in Nijmegen!

Lara and **Xiaochen**, thank you for being my paranymphs. Both of you were there right from the start in 2014 and over the years I've very much appreciated your friendship.

Dank aan onder anderen **Jolanda**, **Vanessa**, **Maaike**, **Femke**, **Claudia** en **Ronny** dat alles altijd zo goed geregeld was bij het DCC. Dankzij jullie, en de rest van het ondersteunend personeel, kon ik me helemaal op mijn proefschrift richten. Hetzelfde geldt voor **Kevin** van het MPI!

There are many other **people at the DCC**, thanks to whom I've always enjoyed the social aspect of the PhD, and whose presence motivated me to come to work. Thank you for making the last 4+ years a wonderful time.

I may not have gone to Nijmegen without the education I received in Oxford. Thank you to **Prof. Hulstijn** and **Prof. Levelt** for helping me transition from Amsterdam to Oxford, and thank you to **Prof. Lahiri** and **Dr. Husband** for helping me transition from Oxford to Nijmegen.

Mijn ouders, **Dorret** en **Frederik**, zijn van het begin tot het einde zeer betrokken geweest. Mijn vader bracht me direct met de auto naar Nijmegen toen ik er op de ochtend van mijn sollicitatiegesprek achter kwam dat de treinen niet reden. En met zijn schilderij op de omslag van dit proefschrift sluiten we de PhD ook weer samen af. De inhoud van het proefschrift is zeker beter geworden dankzij mijn moeder: van het idee om een meta-analyse te doen, tot aan het uitwerken van berekeningen op een kladpapiertje en het lezen van en feedback geven op meerdere hoofdstukken. Dankjulliewel.

Heel veel dank ook aan mijn **familie** en **vrienden** van buiten de universiteit die weliswaar niet direct bij mijn onderzoek betrokken waren, maar wel altijd heel belangrijk voor me zijn.
Pascalle, in de afgelopen jaren heb ik ook jou leren kennen. Dat dit ook de beste jaren van mijn leven zijn geweest, is veel meer dan alleen een correlatie. Zo is de inhoud van mijn lunchtrommel er bijvoorbeeld significant op vooruit gegaan. Maar andere dingen zijn natuurlijk nog veel belangrijker geweest: je warmte en aanwezigheid, de lol en de goede gesprekken, en je begrip en steun. Gelukkig neem ik nu alleen afscheid van mijn promotieonderzoek, maar niet van jou.

CURRICULUM VITAE

Johanna de Vos (Amsterdam, 1989) studied Linguistics and German at the University of Amsterdam, before embarking on a research master's in Linguistics at the University of Oxford. During her bachelors' degrees, she developed an interest in the field of second language acquisition, and during her master's degree she realised her interest in programming and statistics. When the opportunity arose to combine these interests doing a PhD in Nijmegen, she immediately applied. During the following four years, she combined academic research with extra courses in the fields of statistics, data science and artificial intelligence. Since February 2019, Johanna has been working as a software developer at Yoast, a company in Wijchen that focuses on search engine optimisation.

PUBLICATIES

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., & Lemhöfer, K. (2019). *Does study language (Dutch versus English) influence study success of Dutch and German students in the Netherlands?* Manuscript submitted for publication.

De Vos, J. F., Schriefers, H., & Lemhöfer, K. (2019). *Studying in Dutch or English: Does it affect language development?* Manuscript in preparation.

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

De Vos, J. F., Schriefers, H., Ten Bosch, L., & Lemhöfer, K. (2019). Interactive L2 vocabulary acquisition in a lab-based immersion setting. Manuscript accepted for publication.

Koch, E., **De Vos**, J. F., Lemhöfer, K., Housen, A., & Godfroid, A. (2019). *Learning second language morphosyntax in dialogue under explicit and implicit conditions: An experimental study*. Manuscript in preparation.

Wanrooij, K., **De Vos**, J. F., & Boersma, P. (2015). Distributional vowel training may not be effective for Dutch adults. Paper 670. *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, 10-14 August 2015.

DONDERS GRADUATE SCHOOL FOR COGNITIVE NEUROSCIENCE

For a successful research Institute, it is vital to train the next generation of young scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School for Cognitive Neuroscience (DGCN), which was officially recognised as a national graduate school in 2009. The Graduate School covers training at both Master's and PhD level and provides an excellent educational context fully aligned with the research programme of the Donders Institute.

The school successfully attracts highly talented national and international students in biology, physics, psycholinguistics, psychology, behavioral science, medicine and related disciplines. Selective admission and assessment centers guarantee the enrolment of the best and most motivated students.

The DGCN tracks the career of PhD graduates carefully. More than 50% of PhD alumni show a continuation in academia with postdoc positions at top institutes worldwide, e.g. Stanford University, University of Oxford, University of Cambridge, UCL London, MPI Leipzig, Hanyang University in South Korea, NTNU Norway, University of Illinois, North Western University, Northeastern University in Boston, ETH Zürich, University of Vienna etc. Positions outside academia spread among the following sectors: specialists in a medical environment, mainly in genetics, geriatrics, psychiatry and neurology. Specialists in a psychological environment, e.g. as specialist in neuropsychology, psychological diagnostics or therapy. Positions in higher education as coordinators or lecturers. A smaller percentage enters business as research consultants, analysts or head of research and development. Fewer graduates stay in a research environment as lab coordinators, technical support or policy advisors. Upcoming possibilities are positions in the IT sector and management position in pharmaceutical industry. In general, the PhDs graduates almost invariably continue with high-quality positions that play an important role in our knowledge economy.

For more information on the DGCN as well as past and upcoming defenses please visit:

http://www.ru.nl/donders/graduate-school/phd/