MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

WWW.MPI.NL

What was that Spanish word again?

Investigations into the cognitive mechanisms underlying foreign language attrition

Anne Mickan

What was that Spanish word again?

Investigations into the cognitive mechanisms underlying foreign language attrition

Anne Mickan

The work described in this thesis was carried out at the Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, and the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, with financial support from the International Max Planck Research School for Language Sciences (fellowship 2016-2020).

The educational component of the doctoral training was provided by the International Max Planck Research School (IMPRS) for Language Sciences. The graduate school is a joint initiative between the Max Planck Institute for Psycholinguistics and two partner institutes at Radboud University – the Centre for Language Studies, and the Donders Institute for Brain, Cognition and Behaviour. The IMPRS curriculum, which is funded by the Max Planck Society for the Advancement of Science, ensures that each member receives interdisciplinary training in the language sciences and develops a well-rounded skill set in preparation for fulfilling careers in academia and beyond.

What was that Spanish word again?
Investigations into the cognitive mechanisms underlying foreign language attrition

Design / Lay-out
Proefschriftenbalie, Nijmegen

Print
Ipskamp Printing, Enschede

ISBN 978-94-92910-23-3

What was that Spanish word again?

Investigations into the cognitive mechanisms underlying foreign language attrition

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 donderdag 11 maart 2021
om 10.30 uur precies

door

Anne Mickan

geboren op 30 november 1990
te Dresden, Duitsland

Promotor
Prof. dr. James M. McQueen

Copromotor
Dr. Kristin M.L. Lemhöfer

Manuscriptcommissie
Prof. dr. Rob Schoonen
Prof. dr. Merel C.J. Keijzer, *Rijksuniversiteit Groningen*
Dr. Gabriele Janzen

Table of Contents

CHAPTER 1

General Introduction

This chapter is partially based on:

Mickan, A., McQueen, J. M., & Lemhöfer, K. (2019). Bridging the gap between second language acquisition research and memory science: the case of foreign language attrition. *Frontiers in Human Neuroscience, 13*, 397. https://doi.org/10.3389/fnhum.2019.00397

1.1 | Introduction

Not too long ago, I received an unexpected call from an old friend. Cristina and I had met during my undergraduate studies in Berlin in 2009. I had just started studying Spanish and she, originally from Spain, was on Erasmus exchange. By the end of our shared time in Berlin, I had not just gained a very good friend, but I had also spent so much time with her and her Spanish friends that I was pretty fluent in Spanish. Today, I no longer speak Spanish regularly, or at all, and so when Cristina unexpectedly called the other day, my lack of practice became painfully apparent. My Spanish conversational skills were about as elaborate as they were back when we first met. I was constantly searching for words and expressions, switching to English more frequently than not, and much to Cristina's confusion, Dutch words kept creeping into my utterances without me even noticing.

Sure, it had been a while since I had last spoken Spanish to anyone, but it was still baffling to see just how much I was struggling in a language that I had once been able to speak so well. I'm not alone with this experience. Anyone who has learned a foreign language and subsequently stopped using it will be familiar with this defeating feeling that all the time and effort once spent on learning the language was in vain. How come we forget foreign languages so easily, and what determines how fast and how much of the language we lose? Is it time and disuse alone that let language skills erode, or are there other processes involved in driving attrition? In this thesis, I seek answers to these questions and hope to contribute to our understanding of the cognitive mechanisms underlying foreign language attrition.

1.1.1 | Defining Foreign Language Attrition

Although interest in the decline of language skills can be traced back a long time, the study of language loss in healthy individuals only became a field of research in its own right in the early 1980s (see Lambert & Freed, 1982, for the first volume dedicated entirely to language attrition). In the 40 years since, the field has made many advances and a myriad of empirical studies on the attrition of both native and foreign languages have been published. Yet we still know much less about the deterioration of language skills than we do about their acquisition.

The term 'language attrition' is most typically understood to refer to non-pathological language loss in healthy individuals, as opposed to language loss after, for example, stroke or in Alzheimer's disease (see Barkat-Defradas et al., 2019). Even non-pathological language loss comes in many facets though. By far the most studied type of attrition is the loss of first language (L1) skills among migrants immersed in a foreign language (FL/L2) environment (L1 attrition; see Köpke & Schmid, 2004; Schmid, 2016; Schmid & Köpke, 2019). Though much less represented, the field also encompasses the study of language decline in the elderly, including both loss of L1 skills in the L1 environment (see Goral, 2004, for an overview) and loss of L2 skills in an L2 environment (i.e., healthy aging migrants who revert to their L1, e.g., de Bot & Clyne, 1989; Schmid & Keijzer, 2009). The studies presented in this thesis are concerned with yet another type of attrition, also much less studied than L1 attrition: the loss of foreign language skills in an L1 environment.

Within this subfield, Schmid & Mehotcheva (2012) proposed a distinction between the loss of naturally acquired foreign languages (L2 attrition) and the decline of foreign language skills acquired primarily in an instructed setting at school or university (FL attrition). Research on the forgetting of naturalistically acquired foreign languages has often investigated children, who had lived in a foreign country but then returned to their home countries, where they reverted to their L1 and showed signs of (sometimes rapid) decline in the L2 (e.g., Cohen, 1989; Kuhberg, 1992; Olshtain, 1986, 1989; Tomiyama, 2000, 2008). Research on the loss of instructed foreign language skills, in turn, typically deals with (young) adults that learned the foreign language at school or university, but subsequently stopped using it (e.g., Bahrick, 1984a,b; Engstler, 2012; Gardner et al., 1987; Grendel, 1993; Mehotcheva, 2010; Wang, 2010; Weltens, 1988). The majority of experiments conducted within the scope of this thesis (Chapters 2 to 4) come closest to modelling this latter type of attrition, that is the forgetting of foreign languages learned later in life in an instructed rather than natural setting, henceforth simply referred to as foreign language (FL) attrition. Chapter 5, instead, studies the attrition of foreign language knowledge acquired under both instructed and natural contexts in a group of university students that return from a study abroad.

1.1.2 | What Do We Know About Foreign Language Attrition?

1.1.2.1 | *What Do We Forget?*

Foreign language forgetting can be observed on all linguistic levels. It often most clearly manifests in a decrease in verbal fluency and lexical diversity (e.g., Bahrick, 1984a,b; Mehotcheva, 2010), but can also affect grammar (i.e., morphosyntax; e.g., Bahrick, 1984a,b; Weltens, 1988) and pronunciation (e.g., Dugas, 1999). Vocabulary has sometimes been claimed to be most vulnerable to forgetting, but there is no consistent empirical support for the claim that we are more likely to forget the words of a foreign language than its grammar or pronunciation rules (compare Kuhberg, 1992; Moorcroft & Gardner, 1987; Weltens, 1988; see Yoshitomi, 1999, for a discussion of whether it makes sense at all to compare forgetting across different linguistic domains). Nevertheless, lexical knowledge has received somewhat more attention than grammar and pronunciation, especially in recent studies on FL attrition. The studies reported on in this thesis also concern lexical rather than syntactic or phonological FL attrition.

For vocabulary, it appears that cognates (i.e., form-similar words across languages, e.g., English 'table' and Dutch 'tafel') and concrete words are more resilient to forgetting than non-cognates and abstract words (e.g., de Groot & Keijzer, 2000; Weltens, 1988). Some research additionally suggests that low frequency words might be forgotten faster than high frequency words (e.g., Mehotcheva, 2010; though see de Groot & Keijzer, 2000). As for the order of forgetting, there is some evidence that we first forget the foreign words and grammatical structures we learned last (or possibly those learned least well; e.g., Kuhberg, 1992; Hedgcock, 1991; Olshtain, 1989; for more information on this so-called Regression Hypothesis see Chapter 5). Furthermore, it has been established that productive skills deteriorate faster than receptive skills (e.g., Bahrick, 1984a,b; de Groot & Keijzer, 2000). An attriter may be unable to freely recall and produce a word, but still recognize and understand it when they read or hear the word. In earlier studies on FL attrition, the focus was often on receptive language knowledge, as assessed, for example, in multiple choice vocabulary tests or listening and reading comprehension tests. Perhaps not surprisingly so, some of those studies reported very little to no attrition even after years of no exposure (e.g., Murtagh, 2003; Weltens, 1988), while studies with productive recall tasks tended to report significant attrition already within the first year of disuse (e.g., Bahrick, 1984a,b; Mehotcheva, 2010; though see Engstler, 2012). While this is not the only reason for differences in the amount of attrition reported in earlier studies, the way in which attrition is measured matters and needs to be taken into account in interpreting study outcomes.

1.1.2.2 | *How Fast Do We Forget?*

Based on the existing FL attrition literature, forgetting seems to set in rather quickly after one stops using a foreign language and appears to be most severe in the initial years of disuse (e.g., Bahrick, 1984a,b; Wang, 2010). In his seminal study on the retention of school-learned Spanish, for example, Bahrick (1984a,b) found that most forgetting happened in the first three to six years of no exposure to Spanish. Attrition rates subsequently leveled off, with little to no extra loss incurred in subsequent years of disuse. Most of the knowledge that was not lost within the initial attrition period, in fact, remained accessible for another 25 years, leading Bahrick to propose that some foreign language knowledge is stored in what he called 'permastore', where it is less susceptible to forgetting and hence remains intact despite little or no exposure to the target FL.

While it might seem that time is the single most important predictor of forgetting, research has shown that the length of the attrition period alone (i.e., the time since active use of the FL stopped) often does not suffice to explain forgetting (e.g., Mehotcheva, 2010; Murtagh, 2003; Taura, 2008). Exactly how severe and fast the attrition process is, in fact, varies from person to person and appears to depend on a number of factors (for a detailed discussion of individual differences in FL attrition, see Mehotcheva & Mytara, 2019, and Chapter 5). By far the most reliable predictor of forgetting rates appears to be the level of proficiency reached in the foreign language prior to attrition onset. Bahrick (1984a,b), for example, showed that individuals who had taken more Spanish courses and who had obtained higher course grades by the end of training forgot relatively less than participants who had reached lower levels of proficiency (see also Weltens, 1988). Based on these findings, it has even been proposed that there might be a critical threshold in FL proficiency, which once reached saves the learner from subsequent attrition (Pan & Berko-Gleason, 1986; Weltens, 1988). Relatedly, the amount of experience with the foreign language prior to attrition onset has also been argued to affect forgetting rates (e.g., Hansen, 1999). To the extent that more experience leads to higher proficiency in the foreign language, this is not surprising. Yet, the amount of exposure to a foreign language does not take the quality of the input into account, and hence does not necessarily lead to higher proficiency, which might be why the empirical support for an effect of amount of experience on forgetting rates is not as consistent as that of the attained proficiency level (e.g., Mehotcheva, 2010).

Next to these two aspects, one would assume that language use is another crucial determinant of attrition severity: continued, even just occasional use of a foreign language after returning from a study abroad, for example, should counteract attrition. As I discuss in more detail in Chapter 5, however, previous research has paradoxically often failed to establish a beneficial role of language use for target language retention, or conversely a detrimental role of non-target language use for target language retention (e.g., Bahrick, 1984a,b; Mehotcheva, 2010; but see Alharthi & Al Fraidan, 2016). A similar story can be told about the impact of motivational and attitudinal factors, which likewise have not consistently been linked to forgetting / retention (compare Wang, 2010, with Mehotcheva, 2010; Xu, 2010; see Chapter 5 for a more detailed discussion).

Finally, age at attrition onset and literacy also seem to play a role in determining the severity of attrition (see Mehotcheva & Mytara, 2019, for an elaborate discussion). In child returnees, for example, it has sometimes been found that children who leave the foreign language environment aged 8 years or older tend to attrite less than children who return to the L1 environment long before they reach puberty (e.g., Cohen, 1989; Olshtain, 1986; but see Kuhberg, 1992 and Reetz-Kurashige, 1999). Older children might have had more exposure to the foreign language than younger ones, and hence might have learned the foreign language better and attrite less because of that (see proficiency discussion above). Age, however, is also confounded with brain maturation and the development of literacy in children, the latter of which has separately been shown to influence attrition. Adult English native speakers who mastered the Japanese writing system, for instance, subsequently maintained their spoken Japanese better than those who did not master the writing system (Hansen & Newbold, 2001; see Hansen, 2001, for a summary of similar findings with other languages). It is thus unclear what drives the age effects in child attriters. Finally, age might play a role in elderly FL users (see Higby et al., 2019, for L1 attrition; de Bot & Clyne, 1989 for FL decline in L2 environments). For the research reported in this thesis, neither literacy nor age are particularly relevant though since all our participants were aged between 18 and 35 and were literate in all languages they spoke, including the foreign language under investigation.

Clearly, FL attrition is a complex phenomenon. Many of the variables that have been found to be predictive of forgetting are interrelated and more research is necessary to disentangle the relative contributions of each of them to foreign language attrition. Regardless of these individual differences though, it is interesting to note that in most cases foreign language attrition appears to be temporary and reversible. Studies using the so-called 'savings paradigm' (initially developed by Ebbinghaus, 1885, 1913) have demonstrated that relearning seemingly lost foreign language vocabulary is much easier and faster than learning new foreign language vocabulary from scratch (e.g., de Bot et al., 2004; Hansen et al., 2002). Such relearning advantages indicate residual storage of the purportedly 'forgotten' words, and thus speak against complete loss of memory traces. Foreign language attrition is thus best understood as a performance problem, characterized by accessibility difficulties rather than actual loss (Sharwood Smith, 1989). This view of forgetting guides the experimental work in this thesis and also aligns well with how forgetting is defined in the domain-general memory literature, which as will become apparent in the subsequent sections, is central to this thesis.

1.1.2.3 | *So Why Do We Forget?*

While the studies discussed above have added a great deal to our understanding of how fast we forget and what type of foreign language knowledge we lose first, we are still far from a cognitive theory of foreign language attrition and from understanding *why* it happens in the first place. One reason may be the sheer complexity of the phenomenon. As the previous section illustrates, foreign language forgetting is inherently variable and difficult to generalize. Another reason, however, might pertain to how the study of FL attrition is typically approached. Although recent years have seen a surge of interest in more psycholinguistically inspired approaches to FL attrition (e.g., Paradis' Activation Threshold hypothesis, Paradis, 1993; see 1.1.6 for details), most research to date remains documentary in nature. The difficulty with observational, documentary studies is that they can only measure the outcome of the attrition process, such as the inability to retrieve certain words. They offer, however, no way of ascertaining what events caused the observed retrieval difficulties. Of course, researchers can speculate by drawing links between background variables, such as language use or literacy, and the degree of loss; yet these links are purely correlational and do not necessarily entail causality. To further our understanding of the processes *driving* FL attrition, and thus to better understand why we forget (words from) foreign languages, the majority of experiments presented in this thesis (Chapters 2-4) take a different approach. Inspired by the memory literature on forgetting, I tried to induce forgetting in the lab rather than observe it under natural conditions. In doing so, I aimed at establishing the conditions that successfully lead to FL forgetting in the lab as compared to those that don't, and hence hoped to define some of the mechanisms that might (at least in part) underlie foreign language attrition in the wild.

1.1.3 | Forgetting in Other Domains and How It Might Relate to Language Forgetting

Foreign language vocabulary is not the only knowledge we lose over time. We also forget where we park our bike, what we needed so desperately from the grocery store, or what that distant friend's name was. Forgetting is part of everyday life and because of that has actually received quite some attention from cognitive psychologists. Research on forgetting dates back to the 19th century and Ebbinghaus' research on the ease of learning and relearning nonsense-syllable sequences (Ebbinghaus, 1885, 1913). Ebbinghaus discovered that relearning seemingly forgotten syllable sequences was faster than learning them initially. He reasoned that the time it took to relearn material reflected the material's memory strength, and he argued that this measure was a much more accurate reflection of recall ability than dichotomous (remembered vs. not remembered) scores. Forgetting, when defined as a decrease in memory strength over time, is then measurable as an increase in the time needed for relearning, and consequently a decrease in what he called 'savings' (i.e., the time saved during relearning relative to the time expended for original learning).

Comparing those relearning savings after different time intervals (ranging from a few minutes up to 31 days), Ebbinghaus found that memory loss was not linear over time, but logarithmic instead: most forgetting (i.e., the steepest drops in savings) occurred over the first minutes and hours after learning and then gradually leveled off. Note that this resembles what Bahrick (1984a,b) observed for the retention of school-learned Spanish, just on a smaller time scale. Although the experiments in this thesis do not use Ebbinghaus's method of savings, they do assume, like Ebbinghaus, that forgetting is a continuous rather than an all-or-none process in which items in memory are either fully retained or completely forgotten (I return to this in Chapter 6). Ebbinghaus' research kick-started the experimental study of forgetting and inspired a number of theories on why forgetting happens (for an overview, see Anderson, 2015; Ecke, 2004). The most influential of these is interference theory.

1.1.4 | Forgetting Due to Competition and Interference: A Domain-General Perspective

Rather than assuming that forgetting is a by-product of time (cf. decay theory, Thorndike, 1914), interference theory attributes forgetting to competition from other memories. Such competition can stem from 'old' memories that interfere with the acquisition of new knowledge (i.e., proactive interference: Keppel & Underwood, 1962) or with the retrieval of other 'old' memories (i.e., retrieval-induced forgetting, RIF: Anderson et al., 1994; see Chapters 2 and 3), but competition can also come about through the formation of new memories that interfere with the retention of already existing knowledge (i.e., retroactive interference: Müller & Pilzecker, 1900; see Chapter 4). In general, interference theory relies on the fact that memories are interconnected and that those that share a common retrieval cue compete with one another for selection upon presentation of that cue. For the bike parking lot situation mentioned earlier, for example, your parking spot from yesterday competes with your memory of today's parking spot, the common cue being your bike and the abstract knowledge of it being parked somewhere. In experimental psychology, the study of interference-based forgetting is usually done through much more abstract material, such as category-exemplar pairs (e.g., banana and apple both being exemplars of the category of FRUITS) or entirely arbitrary word pairings that are only meaningful within the specific experimental context (e.g., tiger-chair, horse-glue).

One example of forgetting by competition is retrieval-induced forgetting (RIF, Anderson et al., 1994). In a typical RIF study, participants first study a number of category-exemplar pairs (e.g., FRUIT-apple, FRUIT-banana, FURNITURE-table). This phase is followed by selective retrieval practice of some exemplars from some of the categories (FRUIT-banana, but not FRUIT-apple or FURNITURE-table). Finally, recall is tested for all originally studied pairs. Recall is naturally best for the practiced pairs (FRUIT-banana), but importantly, it is worse for unpracticed exemplars from practiced categories (FRUIT-apple) than for unpracticed exemplars from unpracticed categories (FURNITURE-table). The mere act of retrieving information can thus hamper access to information related to the practiced material. Forgetting in RIF studies is typically attributed to executive control processes. Competitors during the retrieval practice phase (e.g., apple) are believed to be inhibited, making them harder to retrieve at final test (e.g., Anderson, 2003; Román et al., 2009; though see Raaijmakers & Jakab, 2013; Williams & Zacks, 2001, for alternative explanations). RIF effects have been demonstrated with a wide variety of stimulus materials. It thus appears to be a generalizable phenomenon (for a review, see Storm et al., 2015).

Another example of forgetting through interference are retroactive-interference (RI) effects, which describe forgetting through new learning (rather than mere retrieval as in RIF). RI effects are typically studied in a paradigm that asks participants to learn two lists of associations in immediate succession (list 1 containing A-B pairings and list 2 containing A-C pairings, see Wixted, 2004, for a review). At a final test of list 1, participants are typically found to be much worse at recalling A-B pairings if they learned list 2 compared to when they did not. In the classical design, this interference effect can be explained through the disruption of the consolidation process of list 1 associations. As explained in more detail in Chapter 4, new memories are thought to undergo a slow integration / stabilization process after encoding to enter long-term memory (e.g., McGaugh, 2000; Walker et al., 2003). This consolidation process is disrupted through the learning of list 2. Retroactive interference effects are, however, also observed with already consolidated list 1 material if the learning of list 2 immediately precedes the final test of list 1. In this case, list 2 learning does not affect consolidation of list 1 associations, but instead leads to competition at final test, inducing interference much like in the RIF studies described above (Wixted, 2004). Chapter 4 discusses these differences in more detail. For now, suffice it to say that competition between memories is key to theories of forgetting. Memories, both new and old, are thought to interact with one another by virtue of being connected to shared cues and this competition has long-term consequences for recall ability.

1.1.5 | Competition and Interference in the Language Domain

In this thesis, I ask whether the interference account of forgetting, and RIF and RI more specifically, is applicable to the foreign language attrition context. Support for the idea that this might be the case comes from the assumption that while speaking, and especially when speaking in a foreign language, words in all the languages a speaker knows become co-activated (see Kroll et al., 2006, 2008, for reviews). Describing to my Spanish friend Cristina how I bike to university every morning, for example, supposedly activated not only the Spanish word 'bicicleta', but also the English label for bike, as well as its German and Dutch translation equivalents

‘Fahrrad’ and ‘fiets’. Evidence for the automatic parallel activation of multiple languages comes, for example, from studies that show that bilinguals’ naming performance in one language is influenced by their other language without the task requiring such co-activation. Costa et al. (2000), for instance, found that bilinguals are faster at naming cognates than non-cognates, while monolinguals name both types of words equally fast. In a phoneme monitoring task, in turn, Colomé (2001) found that bilinguals are slower to decide that a phoneme is not present in a target word when it is present in its translation equivalent. In both cases, only one language was relevant for the task and yet the other language of the bilinguals influenced performance (positively in the cognate case and negatively in the phoneme monitoring case) and hence must have been activated to some extent.

Further evidence for language co-activation comes from picture-word interference studies, in which participants have to name pictures in, for example, their L2 while ignoring visual or auditory distractor words in L1 or L2 (e.g., Costa et al., 1999; Hermans et al., 1998). A striking finding from some of these studies is that an L2 distractor word (e.g., ‘bench’) that is phonologically related to the L1 translation (e.g., Dutch ‘berg’) of the to-be-named picture (e.g., English ‘mountain’) interferes with production of the L2 word (compared to an unrelated distractor). This so-called *phono-translation effect* suggests that the L1 translation equivalents were activated and interfered with retrieval of the L2 words, again despite the fact that participants had no reason to activate the L1 and were only naming pictures in their L2 throughout the entire experiment. Paradoxically though, when the distractor was the L1 translation of the L2 word, naming in L2 was facilitated rather than inhibited and hence faster compared to when the distractor was entirely unrelated (Costa et al., 1999; Hermans, 2004). Intuitively, one would assume the direct activation of a translation equivalent (just as its indirect activation through a phono-translation distractor) to result in competition between lexical alternatives, and hence interference rather than (conceptual) facilitation. Since this is not the case, it remains somewhat unclear to what extent and under which circumstances exactly coactivation of translation equivalents results in competition and thus when it is beneficial and when it is detrimental (compare Kroll et al., 2006, and Green, 1998, with Costa & Caramazza, 1999, and Finkbeiner et al., 2006, for a more detailed discussion).

If words in different languages at least sometimes compete for selection, the question arises how bilinguals solve this competition and select the correct language to speak in. To avoid unwanted language selection errors, a common assumption is that speakers inhibit the competing language during speaking (Green, 1998). Such inhibition is believed to be applied on both on the whole language (global) level and on the item (local) level. Most evidence for global language inhibition comes from so-called language switching studies, which have shown that naming a picture in one language slows down subsequent naming of other pictures in a different language (e.g., Meuter & Allport, 1999; Costa & Santesteban, 2004). Supposedly, these *switch costs* reflect that non-target languages compete and hence need to be inhibited to allow for successful retrieval of words in the target language, making it harder to reactivate the suppressed language on subsequent (switch) trials (see Kroll et al., 2008 for a discussion). Relatively few studies have investigated item-level switching effects, that is the effect of naming a picture in L1 or L2 on later naming of the same picture in the other language. In contrast to the facilitatory translation effects mentioned above, the results from these few studies suggest that there are inhibitory links between translation equivalents, and that these might be even stronger than the more global, whole-language inhibition effects (Kleinman & Gollan, 2018; Declerck & Phillip, 2017, but see Branzi et al., 2014, for seemingly contradictory results and Chapter 3 for a discussion of those).

Although clearly more research is necessary, overall, it seems that there is a reasonable amount of evidence that the languages of a multilingual are simultaneously activated during (especially L2) speech production and some evidence for the fact that translation equivalents in different languages compete with one another as a consequence thereof. In terms of competition for retrieval, translation equivalents (sharing the same concept) might thus be similar to pairs of exemplars in RIF studies (sharing a semantic category cue). If this is the case, the question arises whether between-language competition is a driving mechanism behind language forgetting, and thus whether the between-language competition effects sometimes observed during short-term, online processing also have long-term ramifications.

1.1.6 | Between-Language Interference as a Mechanism Behind FL Attrition?

The idea that attrition is the result of competition between languages is actually not entirely new. Sharwood Smith (1989) as well as Seliger and Vago (1991) already noticed how L1 attrition is influenced by the newly acquired foreign language (L2), for instance, in the form of codeswitches to L2 while speaking in L1. This 'cross-linguistic influence hypothesis' (Sharwood Smith, 1989) is also central to more recent approaches to attrition. Its ideas have been formally discussed for L1 attrition, for example, within the context of Paradis' Activation Threshold Hypothesis (ATH, see Gürel, 2004; Köpke, 2002; Paradis, 1993, 2004, 2007). The ATH assumes that all items in the linguistic system, such as words, are interconnected and influence one another (just as category exemplars in RIF). Each item has an activation threshold (AT). Retrieving a word requires that its activation exceed its AT. The AT is lowered after successful retrieval, but is increased again either gradually through disuse or through top-down inhibition during access of other, competing words, such as translation equivalents in other languages. Heightened activation thresholds then lead to retrieval difficulties because more activation is needed to pass them. Ultimately, ATs can be so high that a word can no longer be accessed, and hence is ‘forgotten’ until accessed again (e.g., during re-learning). For Paradis, (L1 lexical) attrition is thus the result of lack of stimulation of certain words, combined with more recent and frequent access of competing translation equivalents. The ATH makes for a compelling explanation of FL attrition, and in fact is very reminiscent of the interference account of forgetting brought forth within the memory literature. It is also plausible in light of the above reviewed evidence from bilingual language production. However, though frequently invoked in discussions of FL attrition (e.g., Mehotcheva & Köpke, 2019), its role in FL attrition has not previously been tested directly or in a systematic manner (e.g., in experiments where it is attempted to induce attrition through between-language interference, as detailed below).

More generally speaking, despite the apparent similarities and a call to acknowledge and exploit the parallels between domain-general forgetting and language attrition (Ecke, 2004; Linck & Kroll, 2019), surprisingly few FL attrition studies make use of memory theories of forgetting. The experiments that make up this thesis address this gap in the literature. Next to using memory theories of forgetting as a framework within which to think about attrition, I argue that an important added value of the memory theories are the experimental paradigms that have been developed to test them. Via snapshots of FL ability at different time points, observational attrition studies document the result of attrition, but they remain blind to the events and processes that lead to the observed changes. The memory approach to the study of forgetting is quite different: rather than observing forgetting under natural circumstances, experimental paradigms such as the RIF paradigm, try to induce and thus simulate forgetting under tightly controlled circumstances. By manipulating the presence or absence of a presumed cause of forgetting (e.g., interference), they can then define the conditions that do and those that do not lead to forgetting, and hence allow for a more direct test of the mechanisms that drive it. Inspired by the memory literature on forgetting, the experiments in Chapters 2 to 4 thus try to induce foreign language forgetting in the lab. More specifically, they put the interference account of forgetting (and consequently also the ideas behind the ATH) to test. Is between-language competition and interference a driving force in foreign language attrition, and thus, can we make participants forget a recently learned foreign language merely by having them use another one?

1.1.7 | Interference-Induced Foreign Language Attrition – A Lab Paradigm

To test whether between-language competition drives FL forgetting, I used an adapted retrieval-induced forgetting (RIF) paradigm (see Figure 1.1).

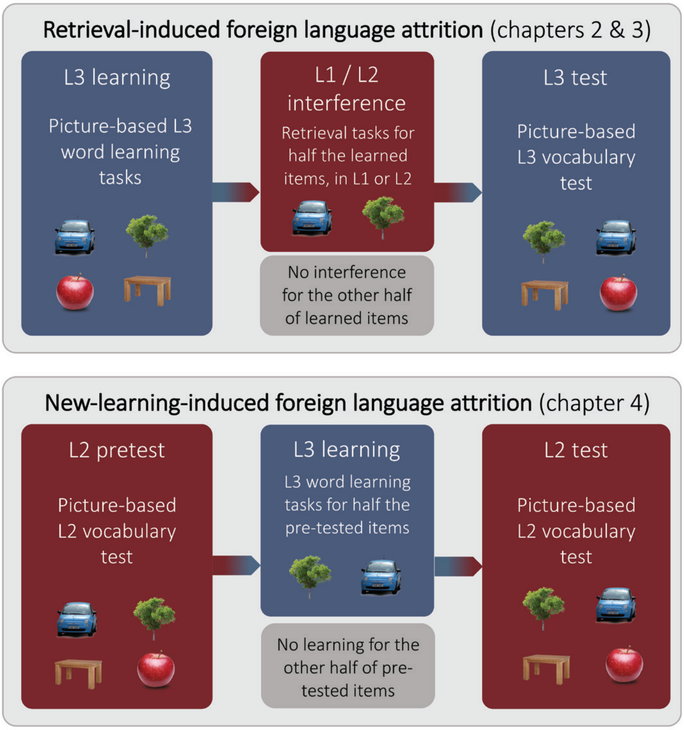
Retrieval-induced foreign language attrition (chapters 2 & 3)
L3 learning
Picture-based L3 word learning tasks
L1 / L2 interference
Retrieval tasks for half the learned items, in L1 or L2
No interference for the other half of learned items
L3 test
Picture-based L3 vocabulary test
New-learning-induced foreign language attrition (chapter 4)
L2 pretest
Picture-based L2 vocabulary test
L3 learning
L3 word learning tasks for half the pre-tested items
No learning for the other half of pre-tested items
L2 test
Picture-based L2 vocabulary test

FIGURE 1.1

Schematic representation of the task design in Chapters 2, 3 and 4.

In Chapters 2 and 3, participants first learned new words in a foreign language (L3 Spanish or Italian). On a subsequent day, I tried to make them forget the recently learned words. According to interference theory and RIF more specifically, one way of interfering with their memory for these recently learned foreign language words should be through the mere retrieval of related memories, in the language case, translation equivalents in other languages. On a subsequent day, our participants were thus asked to retrieve half of the previously learned words in a different language. In Chapter 2, participants retrieved words in either L1 Dutch or L2 English (manipulated between participants). In Chapter 3, interference was induced only through L2 English retrieval. Finally, I assessed what this interference induction did to our participants' ability to retrieve the originally learned foreign language words. If between-language competition is indeed a driving force in FL attrition, intermittent retrieval of translation equivalents in another language should lead to worsened recall of the same words in the target foreign language. The participants should thus be significantly slower and less accurate at recalling words for which they recently retrieved a translation equivalent than for words which were not interfered with. A similar approach has previously been successfully used to induce L1 attrition (Levy et al., 2007; though see Runnqvist & Costa, 2012, for a failure to replicate) and has, in parallel to the work presented in this thesis, also been applied to FL attrition (Bailey & Newman, 2018). Both of these studies will be discussed and compared to our experiments in more detail in Chapter 2.

In Chapter 4, instead, I turned the paradigm around and asked whether new learning of a foreign language can also lead to forgetting of a long-before learned foreign language (see Figure 1.1). To the extent that this study asks what new learning does to previously learned information, it assesses the role of retroactive interference for FL attrition rather than that of interference through retrieval. The fact that the language I tried to interfere with in Chapter 4, however, was learned not within the experiment but in fact long before in real life, sets this study apart from typical retroactive interference studies (as well as from a prior study on RI effects in language attrition: Isurin & McDonald, 2001, discussed in detail in Chapters 2 and 4). The differences and similarities between our experimental approach and that of a typical RI study are discussed in more detail in Chapter 4.

1.1.8 | Foreign Language Attrition and the Brain

The lab-based, experimental approach taken here also allows for the investigation of the neural correlates of foreign language attrition. Neuropsychological studies of *foreign* language attrition are to date extremely rare (see only Osterhout et al., 2019), yet their application to the study of FL forgetting can help us understand its underlying cognitive mechanisms and how these unfold over time. For the latter reason specifically, I made use of EEG recordings in Chapter 3. The electroencephalogram (EEG) reflects the synchronized post-synaptic electrical activity of thousands of neurons in the cortex, picked up via small electrodes attached to the scalp. The electrical current generated by these neurons travels rapidly through the brain and the scull. The activity that we measure at the scalp is hence an almost real-time reflection of neural activity and allows us to capture changes in neural processing (irrespective or their spatial origin) with millisecond accuracy. In experimental research, we are particularly interested in the EEG activity in response to certain sensory events, such as the presentation of a to be named picture. By averaging the activity in response to such stimuli over multiple trials, activity that is time- and phase-locked to the onset of the stimulus adds up and random fluctuations in the signal cancel out. The resulting waveform is usually characterized by a series of negative and positive dips, so-called event-related potential (ERP) components. Chapter 3 looked specifically at an early negative deflection, the N2, which is commonly associated with interference and inhibition and is thus a good candidate for a marker of interference-induced foreign language attrition.

Another way of looking at EEG data is by decomposing the signal into its constituent frequency bands before averaging. This is done through a Fourier transform, which estimates the extent to which activity in each frequency band is present over time. This second method picks up on slightly different aspects of the EEG signal: it captures ongoing neural oscillations (rhythmic activity) that are not necessarily phase-locked to the stimulus and hence cannot be detected by the ERP method. Neuronal oscillations in different frequency bands have been linked to different functions. Of particular relevance for Chapter 3 are theta oscillations (4-7 Hz), which have, like the N2, been associated with interference processes, for example, in episodic memory retrieval.

Using both of these EEG analysis techniques, Chapter 3 sought corroborating neural evidence in favor of the idea that between-language competition is a driving force in foreign language attrition. In an adapted RIF paradigm (see Figure 1.1), we tracked not only the neural correlates of retrieval difficulty after attrition had already occurred (i.e., at final test), but also what happened during the preceding interference phase, when forgetting was supposedly induced. Showing a direct relationship between neural activity during the interference phase and later recall ease or ability would provide additional evidence in favor of the idea that forgetting can be the direct consequence of the more recent use of other languages.

1.1.9 | Integrating Lab-Based and Observational FL Attrition Studies

Despite the advantages discussed above, the lab approach is by no means meant to replace traditional observational attrition studies. Attrition is a complex, multi-faceted phenomenon, which is highly simplified in tightly controlled lab environments. For a thorough understanding of FL forgetting, the two approaches will need to complement each other. For that reason, Chapter 5 of this thesis presents a study with real attriters. In a longitudinal design, I followed a large group of German learners of Spanish throughout their study abroad in Spain as well as throughout the first six months back in Germany. In contrast to predictions based on the ATH and the memory literature on interference-induced forgetting, the existing evidence for a role of target and non-target language use in language maintenance in natural attriters is sparse. Is the importance of language use much smaller in real life than theory would have us believe, or have previous studies just failed to measure it adequately? The longitudinal set-up and monthly, percentage-based frequency of use measurements set this study apart from the majority of previous observational studies on language attrition, as discussed in detail in Chapter 5. Finally, to acknowledge the complex nature of the attrition process, I also asked what other factors besides target and non-target language use influence individual forgetting rates.

1.1.10 | Outline of this Thesis

The overarching goal of this thesis is to understand why we forget foreign languages, and more specifically why we lose access to foreign language vocabulary. While FL forgetting is typically studied 'in the wild' in populations of real attriters, I decided to approach the phenomenon from a different angle. Inspired by how the domain-general memory literature investigates forgetting, the first three empirical chapters (**Chapters 2-4**) try to simulate FL attrition in the lab rather than observe it in real life. The simulation approach has the advantage that it allows to study and manipulate the *process* of attrition as it unfolds, which is impossible in purely observational research. By establishing which experimental manipulations successfully induce foreign language vocabulary forgetting and which do not, I hoped to further our understanding of the cognitive mechanisms underlying foreign language lexical attrition.

To that end, **Chapters 2 and 3** use an adapted retrieval-induced forgetting paradigm (see Figure 1.1). The original paradigm was developed to test the interference account of forgetting, that is the idea that forgetting is the consequence of competition from related memories. In the language context, this translates to competition between translation-equivalents in different languages. If between-language interference is a driving force behind FL attrition, it should be possible to induce long-term retrieval difficulties for words in a recently learned language through the mere retrieval of their translation equivalents in other languages. **Chapter 2** asks whether this is the case and whether it matters which language is intermittently retrieved, comparing interference from the native language with interference from another foreign language. Moreover, in a second experiment, **Chapter 2** additionally explores whether differences in interference strength between native and non-native languages are related to frequency of use and hence accessibility differences between them.

Chapter 3 builds on the results of **Chapter 2** and asks what the neural correlates of such interference-based language forgetting are. Using a similar design, **Chapter 3** seeks corroborating neural evidence in favor of the idea that between-language competition underlies the behavioral forgetting effects observed in **Chapter 2**, and furthermore asks when in time these processes act and contribute to forgetting.

Together, **Chapters 2 and 3** explore how the mere use of other languages impacts a recently learned foreign language. **Chapter 4**, in turn, asks whether a long-known foreign language can be interfered with through the learning of a new foreign language and, if so, whether such interference effects emerge immediately during learning or take time to evolve. Newly learned foreign language words might need to be consolidated and integrated into the mental lexicon before they can interfere with their translation equivalents. This idea is explored in the second experiment presented in **Chapter 4**. This chapter is different from the two previous ones in that it tests a different interference mechanism (new learning rather than retrieval of words in a known language) and in that it assesses how a long-ago (rather than recently) learned foreign language fairs under interference induction.

Having explored different possible mechanisms of FL attrition in the lab under tightly controlled conditions, **Chapter 5** returns to a more naturalistic setting and tests the lab findings in a population of real attriters. In a large-scale longitudinal study with L1 German learners of L3 Spanish, **Chapter 5** asked whether it was possible to document the beneficial role of Spanish use for Spanish maintenance, and conversely whether we could find evidence for a trade-off in accessibility between L3 Spanish on the one hand and L1 German and L2 English on the other hand. Do performance decreases in L3 Spanish go hand in hand with fluency increases in L1 German and L2 English? Finally, **Chapter 5** also asked what other individual factors impact language retention after a study abroad, touching upon some of the factors discussed at the beginning of this chapter.

Finally, **Chapter 6** summarizes the results from the empirical studies reported on in **Chapters 2 to 5**. This summary is followed by a discussion of the theoretical implications of these results and an outline of potential avenues for future research.

CHAPTER 2

Between-Language Competition as a Driving Force in Foreign Language Attrition

This chapter is based on:

Mickan, A., McQueen, J. M., & Lemhöfer, K. (2020). Between-language competition as a driving force in foreign language attrition. *Cognition, 198*, 104218. https://doi.org/10.1016/j.cognition.2020.104218

I would like to thank Laurel Brehm and Johanna de Vos for statistical advice, as well as Alicia Abad Perales and Aurora Troncoso Ruiz for the Spanish audio recordings. I also thank Julia Egger and Iris Verpaalen for help with data acquisition and Michael Masson and Marc Brysbaert for helpful comments on an earlier version of this manuscript.

Abstract

Research in the domain of memory suggests that forgetting is primarily driven by interference and competition from other, related memories. Here we ask whether similar dynamics are at play in foreign language (FL) attrition. We tested whether interference from translation equivalents in other, more recently used languages causes subsequent retrieval failure in L3. In Experiment 1, we investigated whether interference from the native language (L1) and/or from another foreign language (L2) affected L3 vocabulary retention. On day 1, Dutch native speakers learned 40 new Spanish (L3) words. On day 2, they performed a number of retrieval tasks in either Dutch (L1) or English (L2) on half of these words, and then memory for all items was tested again in L3 Spanish. Recall in Spanish was slower and less complete for words that received interference than for words that did not. In naming speed, this effect was larger for L2 compared to L1 interference. Experiment 2 replicated the interference effect and asked if the language difference can be explained by frequency of use differences between native- and non-native languages. Overall, these findings suggest that competition from more recently used languages, and especially other foreign languages, is a driving force behind FL attrition.

2.1 | Introduction

While we have come to understand quite a lot about how we *learn* foreign languages, very little is known about what happens to a foreign language when we no longer use it regularly. If you have ever learned a foreign language, you surely have experienced the very frustrating feeling of not being able to recall the foreign words that just a while back would come to you so easily. Why are you having such a hard time remembering the once so arduously learned words? How come we forget a language's vocabulary so easily?

Research into the forgetting or 'attrition' of languages has to date mostly focused on first language (L1) attrition, the forgetting of one's mother tongue when immersed in a second language (e.g., Choi et al., 2017; Isurin, 2000; Pallier et al., 2003; Pierce et al., 2014, for reviews, see Köpke & Schmid, 2004; Schmid, 2016; Schmid & Köpke, 2019). For foreign language (FL) attrition, only a handful of studies exist, all of which are of a mainly descriptive nature (e.g., Bahrick, 1984a,b; Bahrick & Phelphs, 1987; de Bot & Weltens, 1995; Grendel, 1993; Mehotcheva, 2010; Murtagh, 2003; Tomiyama, 2008; Weltens, 1988; Weltens et al., 1989; Xu, 2010; for an overview, see Mehotcheva & Köpke, 2019; Schmid & Mehotcheva, 2012). A seminal case study by Bahrick (1984a,b) on the retention of school-learned Spanish, for example, showed that most foreign language forgetting happens in the first three to six years and then levels off, with the most basic vocabulary apparently preserved in what he called 'permastore'. Other studies have established that productive skills, as compared to receptive skills, are most vulnerable to forgetting (e.g., Bahrick, 1984a,b; de Groot and Keijzer, 2000), and that we tend to lose first the words and structures we learned last, or possibly those learned least well (also known as the Regression Hypothesis; e.g., Kuhberg, 1992; Olshtain, 1989).

Apart from those general trends, people differ vastly in how much and how fast they attrite in a foreign language. Some studies have failed to observe attrition even after years of no exposure to the foreign language (e.g., Grendel, 1993; Murtagh, 2003; Weltens, 1988), while others find notable forgetting already after just a year or even less (e.g., Bahrick, 1984a,b; Mehotcheva, 2010). Some of those differences can be accounted for by the tests that were used to elicit measures of language retention (e.g., productive vs. receptive tests), however, there are also several study-external

[1] Schmid and Mehotcheva (2012) distinguish between *foreign* and *second* language attrition, the former referring to the attrition of a classroom-learned foreign language and the latter referring to the attrition of naturalistically learned foreign languages. In what follows we use the term foreign language attrition to refer to both types of attrition scenarios.

factors that are believed to impact individual forgetting rates. Among those are proficiency at attrition onset (e.g., Bahrick, 1984a,b; Weltens, 1988; Mehotcheva, 2010), age at attrition onset (e.g., Cohen, 1989; Olshtain, 1986), as well as language usage patterns and motivation to maintain the foreign language (e.g., Mehotcheva, 2010; Wang, 2010). For a recent discussion of extra-linguistic factors in FL attrition, see Mehotcheva & Mytara (2019).

Interestingly, attrition, regardless of its course and rate, has often been found to be temporary. Most convincing evidence of this fact comes from studies that have shown that relearning a FL is a lot easier and faster than learning a new language (e.g., de Bot et al., 2004). Attrition is thus best described as an access problem at the moment of retrieval, rather than actual structural loss. But research has yet to unravel the driving forces underlying this forgetting process. That is, which cognitive mechanisms are responsible for FL attrition?

To approach this question, we took inspiration from the domain-general memory literature (see also Ecke, 2004, for a review of how memory theories of forgetting might be applied to language attrition). After all, forgetting is a very pervasive phenomenon, and by no means unique to the foreign language context. The earliest endeavors to understand forgetting date back to Ebbinghaus and his experiments on the retention of nonsense syllables (Ebbinghaus, 1885, 1913). His work resulted in the famous and still highly influential forgetting curve, which describes the exponential decay of information in memory: most forgetting happens within the first minutes to hours after learning, and then levels off. Inspired by Ebbinghaus' work, Thorndike (1913) later formulated the so-called 'decay theory', which attributes memory loss to neuronal trace decay over time if information is not used and reinforced at regular intervals. His theory, however, received a lot of criticism given that it is virtually impossible to physically observe such trace decay (McGeoch, 1932a). What is more, to provide convincing evidence for decay theory, one would need to show that forgetting happens in the *absence of other events*. Without such proof, one can instead explain memory loss via the occurrence of interfering events, such as the perpetual formation of new memories with time.

In line with the latter notion, interference theory emerged. Rather than attributing forgetting to the mere passage of time and disuse of information, interference theory assumes that forgetting is the consequence of competition between memories. This competition can occur through the formation of new memories that interfere with the retention of already existing knowledge (i.e., retroactive interference: Müller & Pilzecker, 1900; McGeoch, 1932b). Competition can also come from 'old' memories that interfere either with the acquisition of new knowledge (i.e., proactive interference: Keppel & Underwood, 1962) or with the retrieval of other 'old' memories (Anderson et al., 1994). Generally speaking, interference theory relies on the fact that memories that share a common retrieval cue (e.g., semantic category membership, 'banana' and 'apple' both being exemplars of the category FRUIT) compete with one another for selection upon presentation of that cue, and thus interfere with the recollection and retrieval of each other (Roediger, 1973).

One example of forgetting by competition is the retrieval-induced forgetting (RIF) phenomenon. In a typical RIF paradigm, as introduced by Anderson et al. (1994), participants initially study category-exemplar pairs (e.g., FRUIT-apple, FRUIT-banana, ANIMAL-cat), after which half of the exemplars from half of the categories (FRUIT-apple, but not FRUIT-banana, and nothing from the ANIMAL category) are practiced repeatedly. In a final test, all initially studied category-exemplar pairs are tested again. Not surprisingly, recall performance is highest for the repeatedly practiced pairs (FRUIT-apple). Most importantly though, recall for unpracticed exemplars from the practiced category (FRUIT-banana) is worse than recall for exemplars from an unpracticed category (ANIMAL-cat). Retrieving information can thus lead to the forgetting of competing, related, but unpracticed information. It is important to note that forgetting in this context, and in fact in most studies on forgetting, is not necessarily understood as total loss, but rather as retrieval failure or (temporary) inaccessibility. It thus follows that the forgetting effects reported in the current paper also do not necessarily imply complete loss of the tested materials.

This RIF effect is typically explained via an inhibitory mechanism that acts on competing information during the retrieval practice phase (on 'banana' while retrieving 'apple' as an exemplar of the category FRUIT), making the inhibited exemplars difficult to retrieve at later test (Anderson, 2003; though see Raaijmakers & Jakab, 2013, for a different explanation). Studies on the neural correlates of RIF support the inhibitory account of RIF (e.g., Johansson et al., 2007; Penolazzi et al., 2014). The RIF effect has been replicated using a great variety of stimulus materials, including not only category-exemplar pairs in verbal lists, but also pictures (Ford et al., 2004; Hulbert & Norman, 2014; Reppa et al., 2017), faces (Wimber et al., 2015), motor sequences (Tempel & Frings, 2013), and event narratives (MacLeod, 2002), thus establishing it as a pervasive and generalizable memory phenomenon (for a review, see Storm et al., 2015).

The question we asked in this study was whether the interference account of forgetting, and more specifically RIF, is applicable to the FL attrition context and the forgetting of FL vocabulary. In line with Linck and Kroll (2019) and Mickan et al. (2019), we argue that category-exemplar pairs in RIF studies share important properties with concept-label pairs in a given language, and that the RIF paradigm might consequently lend itself well to the experimental study of FL attrition. In both cases, the 'subordinate' entries (the exemplars in the RIF case, the labels in the language case) compete for selection when cued with the 'superordinate' (a given category, or a concept). Just as both 'banana' and 'apple' get activated and compete for selection when exemplars from the category FRUIT are cued, so do labels in different languages upon activation of a given concept (e.g., the English word 'apple' and the Spanish word 'manzana' for the concept ‹APPLE›).

A vast array of studies on bilingual word production provides evidence for language co-activation in lexical retrieval (Kroll et al., 2008). The parallel activation of two (or more) languages while speaking can have both positive and negative consequences. Positive effects are reported in the form of faster access to words when primed with form-similar translations (so-called 'cognates') in a different language (Costa et al., 2000). Mostly though, global processing costs are reported, as manifested in, for instance, longer naming latencies in L2 after naming in L1 and vice versa (i.e., switch costs; Costa & Santesteban, 2004), and a general, permanent naming speed and fluency disadvantage in bilinguals, relative to monolinguals, even in L1 (Gollan et al., 2002; Gollan & Silverberg, 2001). Similar to the RIF context explained above, it is assumed that in order to avoid unwanted language selection errors, speakers need to inhibit the non-target language during speaking (Costa et al., 2006; Linck et al., 2008). Although local competition between translation pairs is not universally found (compare Costa et al., 1999, and Hermans et al., 1998; Kleinman & Gollan, 2018; see also Chapter 1) and although it has been argued that the presence of competition effects and thus the need for inhibition might depend on L2 proficiency and relative language dominance (Costa et al., 2006; Van Hell & Tanner, 2012), the bulk of the evidence shows that languages are activated in parallel and hence that translation equivalents *can* compete for selection (for recent reviews see Kroll et al., 2008, and Kroll et al., 2013). Given this parallel between category-exemplar pairs and concept-label pairs, RIF is likely to be relevant for the between-language situation and might thus be one of the mechanisms behind retrieval difficulties (i.e., attrition) in a foreign language.

In the current study, we tested this hypothesis by asking whether the repeated retrieval practice of translation equivalents in another language leads to later retrieval difficulties in a foreign language. We also asked whether it makes a difference if this retrieval practice happens in the dominant mother tongue (L1), or another foreign language (L2). To our knowledge, we are among the first to investigate this in a systematic manner within the FL attrition context. For L1 attrition, Levy et al. (2007) already provided evidence that repeated retrieval practice in Spanish as a foreign language can lead to the decreased accessibility of the same words in L1 English, as measured in error rates. This study suggests that L1 attrition may be a special case of retrieval-induced forgetting. It is worth noting, though, that memory for the L1 was assessed immediately after L2 Spanish retrieval practice, and in a rather indirect manner, via a rhyme generation task. The immediate effect on L1 memory shows that interference effects persist in the short term, but begs the question whether they also persist for longer, that is for several days, or at least for a delay of 20 minutes, as is common in studies on forgetting and long-term memory (Anderson et al., 1994). Moreover, Runnqvist and Costa (2012) were later unable to replicate the Levy et al. findings in an almost identical set-up, casting further doubt on the generalizability of the original results and calling for more research on the usability of the RIF paradigm in language attrition.

In the FL attrition literature, there is preliminary evidence from a study by Isurin and McDonald (2001) that supports the idea that retrieval failure in a foreign language may be due to interference from another foreign language. In their study, monolingual speakers of English learned a list of words in Russian, a new language to them, right after which they learned another list of partially the same words in Hebrew, yet another new language for them. Immediately after learning, they got tested again on the first list in Russian. Recall for the Russian words that were learned in Hebrew was worse than for the words for which no Hebrew translation equivalent was learned. Again though, there was no delay between interference (i.e., the learning of the second list) and the test (of the list learned first), so this study does not provide evidence for what is typically considered long-term memory in studies on RIF and forgetting more generally. Moreover, the fact that both Russian and Hebrew were entirely unknown to participants prior to the experiment takes this study rather far from real-life forgetting. It is rare that two new languages are learned (almost) simultaneously, and not surprising that the learning of the second list interferes with the first, given the immediate nature of the interference and the lack of consolidation of the first list. It thus remains to be seen whether retrieving genuinely “old” information (i.e., words in languages participants already know) rather than learning something new, also leads to forgetting of recently learned foreign language material.

In the present study, we aimed to address the above studies’ shortcomings. We assessed the role of between-language competition in foreign language attrition by means of a modified retrieval-induced forgetting paradigm consisting of three different phases: an L3 Spanish study phase, an interference phase (corresponding to the retrieval practice phase in RIF studies) in which the participants (native speakers of Dutch) were asked to retrieve half of the recently learned words in another language, and a final L3 Spanish test phase. We hypothesized that the retrieval of translation equivalents would interfere with the accessibility of the newly learned Spanish labels: recall for words that receive interference should be worse than recall for words that do not receive interference, as measured in higher error rates and / or slower reaction times for interfered compared to not interfered words. Importantly, and differently from the studies mentioned above, L3 learning and interference were separated by a night's sleep to allow for consolidation of the newly learned L3 words, and interference and final test were separated by a 20 minute delay (following standard RIF procedure; Anderson et al., 1994) to test for more long-term effects than reported so far. By including another final test one entire week later, we take this last aspect one step further and go beyond traditional RIF studies. If between-language competition is a plausible mechanism for real life attrition, interference effects (although most likely diminished) should persist for a week after interference induction.

Experiment 1 additionally asked whether the source of interference, either the native language (L1 Dutch) or another already known foreign language (L2 English), makes a difference. Given the strength of the L1 and the pervasive evidence for L1 influences on foreign language processing (more so than vice versa; Costa et al., 2000; Gollan et al., 1997), one might expect L1 to be the stronger interferer. However, there is recent evidence suggesting that foreign languages also interfere with one another, and possibly more so than an L1 does with a foreign language. Williams and Hammarberg (1998), for example, report more L2 than L1 influence on L3 productions in a corpus study, and Dewaele (1998) found more L2 than L1 cross-linguistic influence on L3 lexical inventions in another corpus study. A more recent experimental study by Lemhöfer et al. (2020) found L1 and L2 to be equally strong interferers in a picture-word interference paradigm. The interference effect can thus be stronger from either L1 or L2, or can turn out to be equally strong across interfering languages.

2.2 | Experiment 1

2.2.1 | Methods

2.2.1.1 | *Participants*

Fifty-four Dutch native speakers with normal or corrected-to-normal vision and without a history of neurological or language-related impairments were recruited via the Radboud University participant pool. They were randomly assigned to one of two language conditions: interference in L2 English or L1 Dutch. Two participants failed to show up for the second and third sessions of the experiment, for four participants the script failed to construct an appropriate item set (see section 2.2.1.2.1), and five did not reach the learning criterion on the first day (three in the English interference group, two in the Dutch interference group), resulting in a final set of 43 participants (31 female) aged 18 – 34 (*M* = 22.53); there were 23 in the English interference group, and 20 in the Dutch interference group.

Prior to taking part in the study, participants had to fill in an online language background questionnaire ensuring at least some prior experience with Spanish. The amount of experience they had with Spanish ranged from just a few weeks via an online course to a few years of instruction in high school or university. In all cases though, Spanish was the weakest and/or most recently learned foreign language (for frequency of use and proficiency self-ratings, see Table 2.1). We refer to Spanish as an L3, because it was learned *after* L2 English. For some participants Spanish was in fact L4 or even L5; we stick to L3 for simplicity.

For all participants, Dutch was the only native language, and English was the first and most frequently used foreign language. Formal English classes started in elementary school for all participants, though half (N=21) indicated to have had some exposure to English at home before starting school (via video games and TV). Proficiency self-ratings as well as performance on the English LexTALE, a standardized lexical-decision based vocabulary test (Lemhöfer & Broersma, 2012), can be found in Table 2.1. Other foreign languages participants had learned included most prominently French, German and Latin.

The two groups (interference in English or Dutch) did not differ in terms of proficiency or frequency of use self-ratings, age of acquisition or length of exposure in either language, nor did they differ in performance on the English LexTALE or the two executive control filler tasks (see section 2.2.1.3) (all *ps* > .1). The two groups did differ in the amount of time they reported to spend reading in English, with the Dutch interference group reporting higher reading times than the English interference group.

Participants gave informed consent and received either course credit or vouchers for their participation (10€/h). The study was approved by the Ethics Committee of the Faculty of Social Sciences, Radboud University.

TABLE 2.1

Participant characteristics - Experiment 1.

	English interference group (N=23)			Dutch interference group (N=20)		
	M	*SD*	*range*	*M*	*SD*	*range*
	Spanish					
AoA	17.83	3.79	12-26	18.85	4.12	12-29
LoE (months)	21.74	20.25	2-78	14.55	13.84	1 – 60
Frequency of Use (min/day)						
- Speaking	2	5	0-15	9	21	0-75
- Listening	19	33	0-120	37	49	0-180
- Reading	5	29	0-20	15	7	0-120
- Writing	2	5	0-15	9	19	0-60
Proficiency[a]						
- Speaking	1.96	0.93	1-5	2.20	0.83	1-4
- Listening	2.74	1.01	1-6	2.50	0.95	1-4
- Reading	3.00	1.21	1-6	3.10	1.29	1-5
- Writing	1.91	0.95	1-4	2.20	0.95	1-4
	English					
AoA	9.96	2.03	5-13	10.20	1.24	7-12
LoE (years)	11.39	5.21	6-26	11.50	3.40	7-20
Frequency of Use (min/day)						
- Speaking	64	98	0-360	61	114	0-480
- Listening	183	132	59-600	163	94	10-360
- Reading*	83	66	5-240	159	132	10-480
- Writing	56	72	0-300	85	133	0-480
Proficiency[a]						
- Speaking	5.65	0.93	3-7	5.55	0.94	4-7
- Listening	6.17	0.65	5-7	5.95	0.89	4-7
- Reading	6.30	0.70	5-7	6.00	0.65	5-7
- Writing	5.61	1.12	4-7	5.90	0.79	4-7
English LexTALE	77.00	9.21	57-92	76.80	10.47	56-91
	Additional tasks					
Simon task[b]	25	24	-40-61	30	27	-9-102
Go-NoGo FAR	12	9	0-29	8	7	0-40

Note. *M* = mean; *SD* = standard deviation; AoA = Age of acquisition; LoE = length of exposure (i.e., amount of months/years participants had been learning Spanish/English); FAR = false alarm rate (in %). [a] Proficiency self-ratings were given on a scale from 1(very poor) to 7(like a native speaker), [b] The Simon effect is expressed in ms and calculated as the difference between reaction times for the incongruent minus the congruent condition, * indicates that there is a significant difference between groups for this variable.

2.2.1.2 | *Materials*

The stimulus database consisted of 169 Spanish nouns referring to concrete, everyday objects or animals (see Appendix A.1 for the list of words). All these nouns were non-cognates with Dutch and English and were between two and six syllables long in Spanish (M = 2.93), with CELEX frequencies (Dutch lemma frequencies, Baayen et al., 1995) of the Dutch translations ranging from 0 to 72 occurrences per million (M = 13.56). This rather low frequency range was chosen to ensure that there would be enough unknown Spanish words for each participant. For each noun, a photo was chosen from Google images (www.google.com), Flikr (www.flikr.com) or the BOSS database (Brodeur et al., 2010). Photos were all embedded in a 6x6 cm white frame with the depicted object/animal centered and adjusted for size to occupy a maximum of 400 px in either width or length. Furthermore, each noun was recorded by a Spanish native speaker from Madrid (Spain).

2.2.1.2.1 | **Item Selection**

For each individual participant, 40 experimental and 20 filler items were selected on the basis of the participant's pre-test results at the beginning of the first session (day 1), ensuring that the experimental items were all unknown to the participant. When items from the ideal base set (the first 40 items in the pre-test, see 'Pre-test' section under 2.2.1.3.1 for details) were already known to a participant, a Matlab (v.8.6, R2015b, The Math Works, Inc.) script subsequently replaced those items with unknown words from the remaining pre-test items (mean words replaced = 6.16, range = 0-24, see Appendix A.2, for details on the replacement procedure). Each participant's final set of 40 experimental items consisted of two subsets: 20 words that would receive interference on day 2 and 20 that would not. Set assignment to these interference conditions was counterbalanced across participants. Importantly, words in the two subsets were matched on a number of dimensions, including Spanish word length (as measured in syllables), within- and across-set semantic similarity (expressed as a distance value derived from semantic vectors, as described in Mandera et al., 2017), as well as phonological similarity in Spanish assessed via Levenshtein distances (Levenshtein, 1966) (see Table 2.2 for averages). Word frequency was not explicitly controlled for given the amount of constraints we already had. However, as Table 2.2 shows, average frequencies for the subsets were comparable nevertheless. For the interference phase, 20 filler items were chosen in addition to the 20 experimental items that would receive interference. Filler items were not analyzed, and were merely included to disguise the fact that only half of the originally learned experimental items were part of the interference session (for filler item characteristics see Appendix A.3).

Table 2.2
Characteristics of items used in Experiment 1.

	English interference group						Dutch interference group					
	Interference set			No interference set			Interference set			No interference set		
	M	*SD*	*range*	*M*	*SD*	*range*	*M*	*SD*	*range*	*M*	*SD*	*range*
Spanish word length (in syllables)	2.92	0.78	2-5	2.90	0.78	2-5	2.90	0.78	2-5	2.90	0.79	2-5
Dutch log frequency	0.83	0.51	0-1.75	0.87	0.50	0-1.86	0.85	0.51	0-1.86	0.84	0.52	0-1.86
Dutch frequency per million	12.24	13.27	0-56	13.07	13.64	0-72	12.68	14.00	0-72	12.70	14.05	0-72
Within-set semantic distance[a]	0.81	0.11	0-1.07	0.81	0.11	0-1.07	0.81	0.11	0-1.07	0.81	0.10	0-1.09
	M		*SD*		*range*		*M*		*SD*		*range*	
Spanish Levenshtein distance	6.38		1.45		2-12		6.41		1.45		2-12	
Across-set semantic distance[a]	0.82		0.09		0.31-1.09		0.82		0.09		0.46-1.10	

Note. Item sets differed across participants, as described in the Item selection section. Means (*M*) and standard deviations (*SD*) were first calculated per subject and interference condition, and subsequently averaged over groups. Ranges show the absolute min and max values per group and condition.
[a]For an explanation on how we assessed semantic similarity, please refer to Appendix A.2.

2.2.1.3 | *Procedure*

All tasks were administered on a Dell T3610 computer (3,7Ghz Intel Quad Core, 8GB RAM), running Windows 7 and using the stimulus presentation software Presentation (Version 19.0, Neurobehavioral Systems, Inc., Berkeley, CA). The computer screen (BenQ XL 2420Z, 23-inch) was set to white, with a resolution of 1920 x 1080 pixels at a refresh rate of 60 Hz. All audio stimuli were presented to the participants via headphones (Sennheiser HD201), and all oral responses were recorded via a microphone (Shure SM-57) in WAV format using a Behringer X-Air XR18 digital mixer.

Participants were tested individually in a quiet room. They were seated approximately 50 cm from the screen, and about 10 cm away from the microphone. They were told to leave their headphones on at all times during day 1, on days 2 and 8 no headphones were necessary. The experimenter sat in a room next to the participant's room. The door between these two rooms was kept open at all times for efficient communication, and for the experimenter to be able to code the participant's responses (see task descriptions below).

The experiment consisted of three sessions spread out over approximately one week. The general procedure was as follows (see Figure 2.1): On day 1, participants learned 40 new Spanish words in a mixture of recognition and production tasks. One day later, participants were asked to repeatedly retrieve half of these items in either their Dutch or English translation (manipulated between participants). In a final test on day 2, and again roughly one week later (day 8), memory for all 40 items was tested again in Spanish. Cues for naming were always the same pictures, and dependent measures at both final tests included naming accuracy and naming latency.

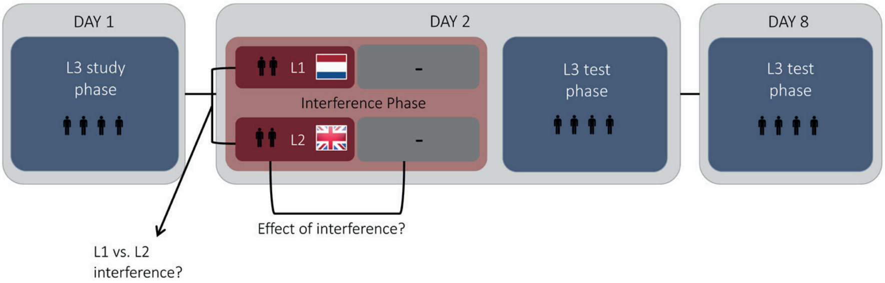

Figure 2.1
Overview of procedure for Experiment 1.

2.2.1.3.1 | Day 1 – L3 Spanish Learning Phase

Pre-test. The first day started with a Spanish picture naming test to select 40 participant-specific, unknown Spanish words for the remainder of the experiment. Participants were asked to name pictures from the database described above in Spanish to the best of their knowledge. They were encouraged to guess when unsure, and to take their time in thinking about the answer. There was no time limit. The experimenter immediately coded participants' answers for correctness. Crucially, the first 40 items of the database constituted the so-called 'base set', the ideal set of items for the experiment, provided they were all unknown to a participant. All participants had to name at least these first 40 pictures. Only if any of those initial 40 words were already known to the participant (which was the case for all but two participants), the remaining (maximum of) 129 pictures had to be named, unknown words out of which would serve as replacement options to re-fill the base set (see Appendix A.2 for details). To make this pre-test as efficient and short as possible, participants who needed few replacements (max. 5) and knew few of the subsequent replacement options, could stop after 101 pictures (N = 14) rather than having to go through all 169 pictures (N = 27).

Whether a word was known in Spanish or not was determined as follows: after a participant had named a picture (or had attempted to do so), they were shown the correct Spanish word on screen. Participants were then asked if they recognized the word. If a participant had not been able to correctly name a picture initially, but recognized the word upon seeing it, it was counted as 'known', and would thus not be used for the experiment. This way, only words that were completely new (i.e., neither named nor recognized by the participant) were included as items in the study.

Learning Tasks. The learning phase consisted of a mix of comprehension and production tasks (see Figure 2.2). The tasks started out easy and got progressively more difficult. With the exception of the final recall test at the end of the learning session (see below), none of the tasks had a time limit. The order of items was semi-random in all tasks, such that it was different for each task and person, but within a task, the order of items was kept constant across rounds. We chose for this type of randomization to avoid order effects in learning, while at the same time keeping distances between repetitions of the same items within a task constant. There were never more than two identical rounds in a row. Feedback was provided on all learning tasks (see details for each task separately). For inter-stimulus intervals and feedback timing please consult Appendix A.4.

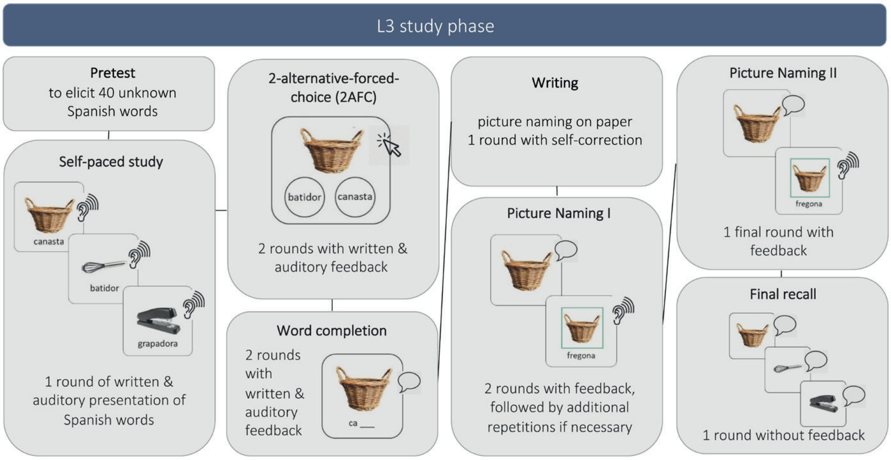
L3 study phase
Pretest
to elicit 40 unknown Spanish words
Self-paced study
canasta
batidor
grapadora
1 round of written & auditory presentation of Spanish words
2-alternative-forced-choice (2AFC)
batidor
canasta
2 rounds with written & auditory feedback
Word completion
2 rounds with written & auditory feedback
ca __
Writing
picture naming on paper
1 round with self-correction
Picture Naming I
fregona
2 rounds with feedback, followed by additional repetitions if necessary
Picture Naming II
fregona
1 final round with feedback
Final recall
1 round without feedback

Figure 2.2
Overview of learning tasks.

The first task (self-paced study) was to familiarize participants with the items. Each Spanish word was presented once in writing together with the corresponding picture and audio. Participants were told to attentively click through all items at their own pace, and were furthermore asked to repeat the audio out loud in order to get used to the pronunciation of the words.

After this initial familiarization phase, participants did two rounds of a two-alternative forced choice task. A picture was presented together with two written words from the stimulus set. Participants were asked to click on the word that was the correct Spanish name for the picture. Upon clicking on a word, they received automatic feedback: the picture remained on screen and either a green (correct) or a red (incorrect) frame appeared around the word they had chosen, and the correct audio was played. In the second round of this task, they were asked to first attempt to name the picture in Spanish before seeing the answer options; otherwise the second round was identical to the first round.

Subsequently, participants did two rounds of a word completion task, in which the picture was accompanied by the first syllable of the Spanish word (for monosyllabic words the first grapheme). Participants had to say the complete word out loud. The experimenter coded their answers for correctness. Only entirely correct productions were counted as correct, but typical Dutch pronunciation errors were not punished and ignored (see Appendix A.5 for details). Based on the experimenter's coding, participants received feedback: again either a red or a green frame around the picture together with the audio and the written form of the correct word.

The word completion task was followed by one round of a writing task. Prompted by a picture, participants wrote down on paper the Spanish word for a picture and then corrected themselves when necessary by rewriting the word on the same piece of paper, this time in red ink, and based on self-initiated feedback on the computer screen (again written and auditory, but without the colored frames).

The last learning task was an adaptive picture naming task. Participants were asked to name the pictures aloud in Spanish, with the experimenter coding the correctness of their responses (same criteria as for the completion task, same feedback). After two initial rounds of picture naming, the words that had not been learned yet received additional exposures. A word was repeated until it had been named correctly (at least) twice in a row. When there were less than ten still-to-be-learned items, already known words reappeared such that one adaptive round contained always at least ten items. Once all words had been learned (i.e., named correctly at least twice in a row) all 40 words were repeated once more to ensure that none of the words had been forgotten over the course of the adaptive learning task.

Finally, participants' knowledge of all 40 Spanish words was tested one last time without feedback to have a final measure of which words were actually learned. Participants had a maximum of 30 s to respond and recordings were made in order to later score their responses for accuracy and naming speed. Naming speed in this final recall test was measured as a baseline for the later naming tasks in Spanish on days 2 and 8.

This learning phase resulted in a minimum of nine exposures per word with feedback, in addition to the final test without feedback. In total, participants thus saw each word a minimum of ten times (*M* = 12.08, mean *SD* = 1.75, abs. range = 10-34). In total, the learning session took maximally two hours. The adaptive learning task was stopped after 1h 45 min, when necessary. Participants had to learn at least 30 out of the 40 words, as measured in the final test, in order to proceed to the next sessions (as noted above, 5 out of 50 participants did not meet this requirement).

2.2.1.3.2 | Day 2 – L1 / L2 Interference Phase & L3 Spanish Post-test

Interference Phase Tasks. One day after the learning session (Dutch group: *M* = 24.03 hours, *SD* = 2.89, range = 18-29, English group: *M* = 24.13 hours, *SD* = 2.41, range = 20.5-32), participants came in for the interference session, which for half of the participants was in Dutch (L1) and for the other half in English (L2). Each participant had to engage with 20 of the initially learned items as well as with 20 filler items nine times in total: once during an initial word completion task, four times during a picture naming task and four times during a letter search task (see Figure 2.3). In the familiarization round, participants saw the picture together with the first syllable of the English or Dutch word (the first grapheme for monosyllabic words) and had to complete the word out loud. After that, they were presented with the correct words on the screen, and were asked to indicate whether they recognized the word. This familiarization round served mostly as a pre-test for the people in the English group to make sure they knew all English words and to take note of those they did not know. As for the pre-test in Spanish on day 1, when a word was not named but recognized, it still counted as known. Completely unknown words in either Dutch (N = 0) or English (total of 14 unknown words for 5 participants; for the entire English group: *M* = 0.61, mean *SD* = 1.37, abs. range = 0-5) were later excluded from analysis.

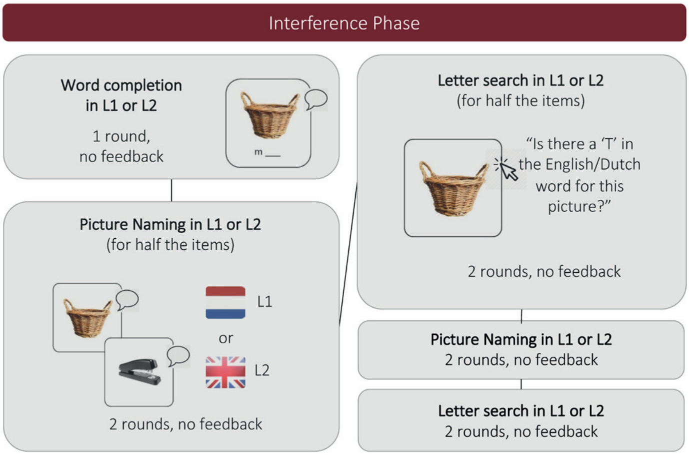
Interference Phase
Word completion in L1 or L2
1 round, no feedback
m__
Letter search in L1 or L2
(for half the items)
"Is there a 'T' in the English/Dutch word for this picture?"
2 rounds, no feedback
Picture Naming in L1 or L2
(for half the items)
L1
or
L2
2 rounds, no feedback
Picture Naming in L1 or L2
2 rounds, no feedback
Letter search in L1 or L2
2 rounds, no feedback

Figure 2.3
Overview of interference phase tasks.

During the subsequent four rounds of standard picture naming (no letter cues presented), no feedback was provided. In the letter search task that followed, participants had to click a button (within 10 s after picture presentation) depending on whether or not the Dutch or English word for the picture contained a certain letter. For each round, participants got a new letter (R,L,T,N). Participants did not get feedback on their performance. The order of items in the interference tasks was semi-randomized. Each task and participant had a different semi-random order, but there were never more than three items from the same condition in one row.

Filler Tasks. Following standard RIF studies (Anderson et al., 1994), after interference and before the final test in Spanish, two distractor tasks were administered to temporally separate these two phases from each other. One of these tasks was the Simon task, the other was the Go-NoGo task (see Appendix A.4 for task design details). Together they took roughly 20 minutes. We chose these specific tasks because they were taxing enough to keep participants from further practicing the recently learned Spanish words, and because they did not require verbal responses, thus creating no additional language interference. Performance of the two groups on these tasks is given in Table 2.1. Since these tasks merely served as filler tasks, we did not analyze them any further.

Final Spanish Test. Finally, and most importantly, in order to assess the effect of the interference phase on Spanish recall, participants were tested on all initially learned items again in Spanish. All pictures were presented in random order, and participants were asked to name them in Spanish. No feedback was provided, and there was no time limit for participants to provide their answers. Accuracy and naming latencies were measured.

2.2.1.3.3 | **Day 8 – Delayed L3 Spanish Test**

About a week after session 1 (English group: *M* = 7.26 days, *SD* = 0.86, range = 6-9; Dutch group: *M* = 6.85 days, *SD* = 0.99, range = 6-9), participants came back for one final Spanish test, identical in format to the final Spanish test on day 2. This session was to test the persistence of the interference effect. During this last session, participants also completed the English version of the LexTALE as a measure of their English vocabulary size (Lemhöfer & Broersma, 2012; see Table 2.1 for group means).

2.2.1.4 | *Accuracy Scoring*

Participants' Spanish word productions (the final utterance in case of multiple attempts) were compared to target (i.e., Spanish native speakers') productions based on phonological similarity (see Appendix A.5 for details). Given that a lot of productions were partially correct (a participant saying 'embuda' for 'embudo'; 13% of all productions and 66% of errors), a binary correct/incorrect scoring was not suitable. Following de Vos et al. (2018), we instead coded responses on the phoneme level and counted the number of correctly and incorrectly produced phonemes for each word. Incorrect phonemes could be either omissions, insertions or substitutions (see Levenshtein, 1966). Table 2.3 exemplifies the scoring procedure for the 'embudo' example.

Table 2.3
Scoring example, phonetically transcribed.

Target word	ε	m	b	u	ð	o
Participants production	ε	m	b	u	ð	a
Scoring	correct	correct	correct	correct	correct	incorrect (substitution)

‘Embuda’ would be counted as having 5 correct phonemes and 1 incorrect phoneme. These two numbers (5,1) form the basis for the dependent variable for statistical modelling (see section 2.2.1.6 for details). For the purpose of providing descriptive statistics and for the figures, we also calculated an error percentage based on these two numbers. This percentage corresponds to the number of incorrect phonemes out of the total number of phonemes, for the ‘embuda’ example: (1/(5+1))*100 = 16.67%.

2.2.1.5 | *Reaction Time Measurements*

Naming latencies were measured manually in Praat (version 5.3.78, Boersma, 2001) from picture presentation until speech onset. Trials on which a participant was unable to name the picture, named it incorrectly or took multiple attempts at naming before succeeding were excluded from analysis. Trials with spill-over from previous trials (the participants correcting themselves), and trials where participants coughed or laughed were also excluded. Smacks, or prolonged thinking sounds (‘uhhh’) were accepted though; naming latencies for these trials were measured at the onset of the actual word production.

2.2.1.6 | *Modelling*

We analyzed the data using generalized mixed effects models with the lme4 package (version 1.1-15, Bates et al., 2015) in R (Version 3.4.3, R Core Team, 2013). Following de Vos et al. (2018), the accuracy data were analyzed using a generalized linear mixed effects model of the binomial family, fitted by maximum likelihood estimation, using the logit link function and the optimizer ‘bobyqa’. A two-column data frame with the number of correct and incorrect phonemes for each target word utterance was passed to the model. Based on these numbers, the model estimated the binomial parameter (i.e., the probability of correctly producing a phoneme for each given word), which was then used for further parameter estimation and hypothesis testing. This 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 Interference (two levels: no interference, interference), Language (two levels: Dutch, English) and Day (two levels: day 2 (immediately after interference), day 8 (one week later)) and their interactions. All fixed effects variables were effects coded (-0.5, 0.5). Random effects were fitted to the maximum structure justified by the experimental design (Barr et al., 2013), which included random intercepts for both Subject and Item, as well as random slopes by Subject for Interference and Day and their interaction. Random slopes were removed when their inclusion resulted in non-convergence to fit the maximum model justified by the data, and when they correlated above 0.94 to avoid over-fitting (Brehm et al., 2019). All p-values were calculated by model comparison, omitting one factor at a time and using chi-square tests.

Naming latencies were analyzed using a linear mixed effects model, fitted by restricted maximum likelihood estimation (using Satterthwaite approximation to degrees of freedom). Naming latencies here refer to the difference in naming latencies between production on day 1 (after learning, serving as baseline) and day 2 and day 8 respectively. These difference scores take into account differences in accessibility that exist between the Spanish words after learning (i.e., due to some words being easier to learn than others). Difference RTs are thus a cleaner measure than raw RTs because they isolate the effect of the interference manipulation on Spanish naming speed at post-test.[2] Raw naming latencies on all days were first log-transformed and then difference scores were calculated and entered into the model. Fixed effects were the same as for the accuracy model and the random effects structure was also determined based on the same principles.

2.2.2 | Results

2.2.2.1 | *Learning Success on Day 1*

Overall, participants did very well on the learning tasks: on average, 95% (*SD* = 5%, range = 78 – 100%) of words were learned. The Dutch and English interference groups did not differ in terms of learning success (Dutch group: 96%, English group: 95%; $t(41) = 0.58$, $p = .565$, $d = 0.177$), or the average number of repetitions needed per item (English group $M = 11.64$, Dutch group $M = 12.60$; $t(41)=1.84$, $p = .073$, $d = 0.563$).

[2] For analyses on the raw naming latencies at final test, rather than difference scores, see Appendix A.6. These analyses lead to the same conclusions as the primary analyses based on difference scores.

2.2.2.2 | *Naming Performance After Interference (on Days 2 and 8)*

2.2.2.2.1 | Naming Accuracy

Figure 2.4 shows the mean percentages of incorrectly recalled phonemes in Spanish on days 2 and 8 per interference and language condition. Words that were not learned on day 1 were excluded from the analysis on a by-participant basis. The percentages given here thus reflect participant-specific proportions: for example, for a participant who learned 36 words, 100% reflects those 36 words rather than the full set of 40 words. Outputs from the mixed effects models are reported in Tables 2.4 and 2.5. We observed a main effect of Interference in line with our predictions: participants indeed made more errors on words that had received interference (12%) compared to words that had not (7%). Similarly, a main effect of Day was observed such that participants made more errors overall a week after interference (12%) compared to immediately after interference (7%). The interference effect was modulated by Day. Separate models fitted for each day (Table 2.5) showed that the interference effect was only significant on day 2 (interfered: 10% errors, not interfered: 4%), but not a week later (interfered: 14%, not interfered: 10%, though numerically still present). There was no main effect of Language, nor any interactions involving this factor.

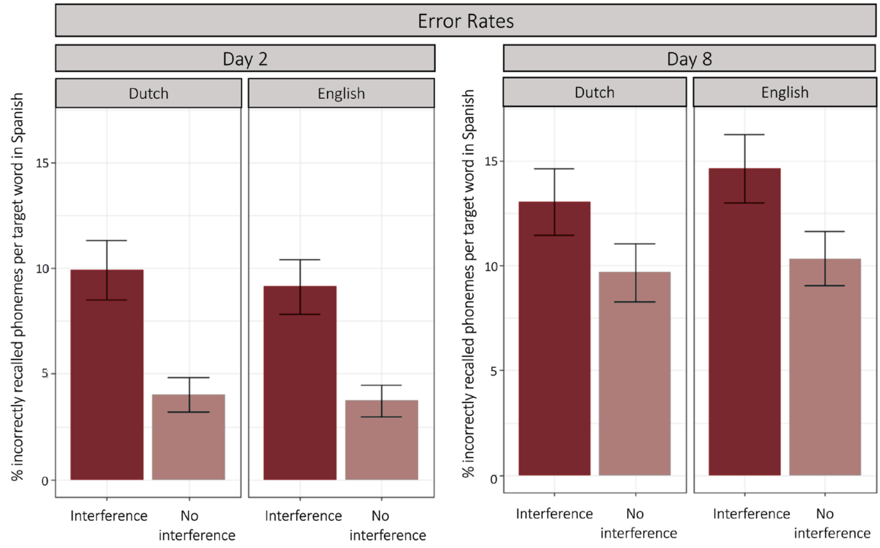

Figure 2.4

Experiment 1. Error rates in Spanish productions as measured in percentage of incorrectly recalled phonemes for the final tests on day 2 and 8 respectively. Error bars reflect the standard error around the condition means.

Table 2.4

Mixed effects model output for naming accuracy in Experiment 1.

Fixed effects	Estimate	*SE*	*z*	*p*(χ^2)		
Intercept	4.06	0.29	13.98	**<.001**		
Interference	-0.58	0.22	-2.58	**.018**		
Language	0.08	0.45	0.17	.865		
Day	-0.64	0.20	-3.26	**.003**		
Interference*Language	0.31	0.43	0.70	.487		
Language*Day	-0.28	0.37	-0.75	.460		
Interference*Day	0.73	0.27	2.67	**.011**		
Interference*Language*Day	0.22	0.48	0.45	.655		
Random effects	**Groups**	**Var**	***SD***	**Corr**		
Item	Intercept	2.34	1.53			
Subject	Intercept	2.07	1.44			
	Interference	1.70	1.30	0.29		
	Day	1.20	1.10	0.33	0.08	
	Int*Day	1.53	1.24	0.21	-0.21	0.06

Note. Significant effects are marked in bold. *SE* = standard error; *p*(χ^2) = Chi-square p-value; Var = variance; *SD* = standard deviation; Corr = correlation.

Table 2.5

Mixed effects model output for naming accuracy split by day in Experiment 1.

	Model output for Day 2				Model output for Day 8			
Fixed effects	**Estimate**	***SE***	***z***	***p*(χ^2)**	**Estimate**	***SE***	***z***	***p*(χ^2)**
Intercept	4.92	0.38	12.96	**<.001**	3.80	0.34	11.10	**<.001**
Interference	-0.98	0.29	-3.43	**.001**	-0.24	0.24	-1.01	.381
Language	0.23	0.43	0.53	.601	-0.11	0.53	-0.22	.823
Interference* Language	0.10	0.53	0.20	.846	0.33	0.45	0.73	.477
Random effects	**Groups**	**Var**	***SD***	**Corr**	**Groups**	**Var**	***SD***	**Corr**
Item	Intercept	4.16	2.04		Intercept	3.07	1.75	
Subject	Intercept	1.76	1.33		Intercept	2.81	1.68	
	Interference	2.26	1.50	0.13	Interference	1.72	1.31	0.33

Note. Significant effects are marked in bold. *SE* = standard error; *p*(χ^2) = Chi-square p-value; Var = variance; *SD* = standard deviation; Corr = correlation.

2.2.2.2.2 | Naming Latencies

Naming latencies (in ms) at final test on day 2 and 8, respectively, split by Interference and Language condition are plotted in Figure 2.5 and model outcomes are shown in Tables 2.6 and 2.7. As indicated above, naming latencies here refer to the difference in naming speed from day 1 (baseline, after learning) and day 2 and 8 respectively, and thus reflect the slowing of responses from baseline to final test. Again, we observed a main effect of Interference indicating that, overall, participants were slowed down more for words that had been interfered with (718 ms) than for words that had not (357 ms). We also again observed a main effect of Day such that participants were overall slower on day 8 (738 ms slower than at baseline) as compared to day 2 (336 ms slower than at baseline).

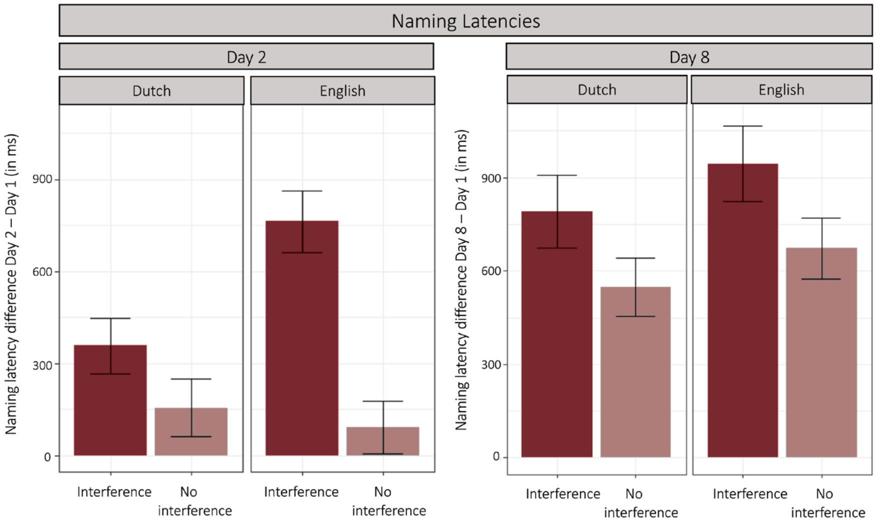

Figure 2.5

Experiment 1. Spanish naming latencies (in ms) at final test on day 2 and 8, respectively. Naming latencies reflect the difference in naming speed between baseline (immediately after learning) and final test (day 2 & 8). Error bars reflect the standard error around the condition means.

The interference effect in naming latencies was modulated by Day such that it was more pronounced on day 2 (interfered: 574 ms, not interfered: 120 ms) than on day 8 (interfered: 872 ms, not interfered: 615 ms), but still statistically significant on both days, as confirmed by follow-up models fit for each day separately (Table 2.7). The interference effect was furthermore modulated by Language such that it was more pronounced in the English group (interfered: 849 ms, not interfered: 365 ms) than in the Dutch group (interfered: 567 ms, not interfered: 346 ms). Finally, we also observed a 3-way interaction among all factors. Follow-up models fit for each day separately showed that the interference effect was modulated by Language on day 2, but not on day 8: the interference effect tended to be more pronounced in the English group (interfered: 763 ms, not interfered: 91 ms, $t(22) = 9.98$, $p < .001$, $d = 2.080$) than in the Dutch group (interfered: 357 ms, not interfered: 155 ms, $t(19) = 3.89$, $p = .001$, $d = 0.870$) on day 2, but this was no longer the case on day 8 (English group: interfered: 944 ms, not interfered: 673 ms, $t(22) = 1.89$, $p = .072$, $d = 0.395$; Dutch group: interfered: 790 ms, not interfered: 548 ms, $t(19) = 2.80$, $p = .012$, $d = 0.625$).

Table 2.6

Mixed effects model output for naming latencies in Experiment 1.

Fixed effects	Estimate	*SE*	*t*	$p(\chi^2)$
Intercept	0.28	0.03	8.73	**<.001**
Interference	0.16	0.02	7.65	**<.001**
Language	0.08	0.06	1.37	.167
Day	0.17	0.03	6.94	**<.001**
Interference*Language	0.11	0.04	2.62	**.009**
Language*Day	-0.01	0.05	-0.26	.782
Interference*Day	-0.14	0.04	-3.25	**.001**
Interference*Language*Day	-0.18	0.09	-2.08	**.037**
Random effects	**Groups**	**Var**	***SD***	**Corr**
Item	Intercept	0.01	0.12	
Subject	Intercept	0.03	0.17	
	Day	0.01	0.08	0.30

Note. Significant effects are marked in bold. *SE* = standard error; $p(\chi^2)$ = Chi-square p-value; Var = variance; *SD* = standard deviation; Corr = correlation.

Table 2.7

Mixed effects model output for naming latencies split by day in Experiment 1.

	Model output for Day 2				Model output for Day 8			
Fixed effects	Estimate	*SE*	*t*	$p(\chi^2)$	Estimate	*SE*	*t*	$p(\chi^2)$
Intercept	0.19	0.03	6.19	**<.001**	0.36	0.04	9.90	**<.001**
Interference	0.23	0.03	8.15	**<.001**	0.10	0.03	2.96	**.003**
Language	0.08	0.06	1.37	.167	0.07	0.07	1.15	.246
Interference*Language	0.21	0.06	3.73	**<.001**	0.02	0.07	0.37	.706
Random effects	**Groups**	**Var**	***SD***		**Groups**	**Var**	***SD***	
Item	Intercept	0.01	0.09		Intercept	0.02	0.13	
Subject	Intercept	0.03	0.17		Intercept	0.03	0.18	

Note. Significant effects are marked in bold. *SE* = standard error; $p(\chi^2)$ = Chi-square p-value; Var = variance; *SD* = standard deviation; Corr = correlation.

2.2.3 | Discussion

In the first experiment, we set out to test the interference account of forgetting in the context of foreign language attrition. More specifically, we asked whether repeated retrieval of words in either the mother tongue or another foreign language would hamper the subsequent retrieval of their translation equivalents in a foreign language, in this case Spanish words that had been recently acquired, but for which there had been an opportunity for overnight consolidation. Experiment 1 showed that this is the case: both in recall accuracy and in recall speed, we observed a disadvantage for recalling Spanish words that had been interfered with compared to those that had not. Moreover, this effect proved to not just be a temporary suppression effect, but persisted for 20 minutes post interference induction and, in reaction times, even for an entire week.

Our results resemble those reported in traditional RIF studies. Those studies established that the repeated retrieval of certain memories (e.g., category-exemplar pairs) interferes with the subsequent retrieval of related, but unpracticed memories (e.g., other exemplars from the practiced category; Anderson et al., 1994; Levy & Anderson, 2002). Our results suggest that similar dynamics can be observed between concept-label pairs. Retrieving a label, a word in for example L1, hampers subsequent access to other labels (i.e., translation equivalents) attached to the same concept. While between-language competition dynamics are well-known to impact language accessibility globally, we confirm that they also act locally, that is between translation equivalents (which had previously sometimes been found to facilitate each other instead, e.g., Costa et al., 1999; see Chapters 1 and 6 for details). What is more, between-language competition had so far mostly been demonstrated during *online* processing and thus in the short term (e.g., switch costs on language switch vs. repeat trials). We show that language competition affects retrieval ease well beyond the single trial, and thus establish it as a phenomenon with long-term ramifications. In doing so, we link competition processes to language attrition, and provide a plausible account of how foreign language forgetting comes about.

We are not the first to draw this link between RIF-like competition processes and language attrition. Our findings are in line with the few prior studies on the topic (Isurin & McDonald, 2001; Levy et al., 2007). These studies, however, focussed on short-term effects of L2 learning on memory for another recently learned L2 (Isurin & McDonald, 2001) or on effects of an L2 on L1 (Levy et al., 2007). Neither of these two studies tested for true long-term effects of interference; their effects are limited to a single experimental session and interference assessment immediately after interference (i.e., without any delay before final test). Moreover, our study adds to these studies in showing that language RIF also applies to consolidated foreign language knowledge, and that the mere retrieval of words from an L1 or another foreign language, as compared to new learning (as in Isurin & McDonald, 2001), is enough to induce forgetting. Taking all these aspects together, our study thus offers a more realistic account of how words are forgotten than earlier studies on retrieval-induced language attrition.

The main effects of interference clearly support our primary hypothesis that between-language interference may be an important factor in driving language attrition. We additionally found the interference effect to differ in magnitude between the two language groups. In naming speed, on day 2, the interference effect was larger for L2 compared to L1 interference. In other words, L3 Spanish recall was more hampered when L2 English interfered than when intermittent retrieval practice took place in the participants' L1 Dutch. While this is surprising given the wide-spread assumption that the dominant L1 should interfere more, this finding is in line with corpus studies reporting a stronger L2 than L1 influence on L3 productions (Williams & Hammarberg, 1998) and L3 lexical inventions (Dewaele, 1998), as well as with studies showing similarly stronger L2 than L1 transfer in the domain of syntax (Bardel & Falk, 2007) and phonology (Llama et al., 2010).

It remains unclear, however, *why* another foreign language would be a stronger interferer than the much more dominant and stronger mother tongue. In Experiment 2, next to replicating the main effect of interference observed in Experiment 1, we aim at providing an answer to this question. One possibility is that the interference difference is inherent to native vs. non-native languages. In the psycholinguistic literature, in response to the above reported studies, it has been argued that foreign languages acquired later in life are grouped together in the mind of the learner and are kept separate from the L1 (Hammarberg, 2001). Such a grouping could explain why foreign languages have sometimes been found to interact more with one another than with the L1. There is, however, to date no corroborating neuroscientific evidence for such a grouping.

Another explanation relates to frequency of use differences between the two languages. One's native language will usually be the most frequently used language in everyday life. An L1 (like Dutch for our participants) thus typically is more strongly represented than any non-native, foreign language (like English in our study). It follows naturally that L2 words are more difficult to retrieve than L1 words. A recent study by Ibrahim et al. (2017) suggests that many processing asymmetries between native and non-native languages boil down to such frequency of use differences.

Why would frequency of use, and resulting ease of retrieval, matter for interference? As briefly touched upon in the introduction (section 2.1), the classic RIF effect is often explained by means of an active inhibitory control mechanism: competing memories during retrieval practice (i.e., during the interference phase) are thought to be inhibited, making these memories more difficult to recall at later test (Anderson, 2003). Applied to the language situation, this means that when retrieving items during the interference phase, in our case words in L1 Dutch or L2 English, related memories, including the recently learned L3 Spanish words, will be co-activated and will be competing for selection, thus hindering the retrieval of the required L1/L2 words. In order to resolve this competition, and to ensure successful retrieval of L1/L2 words, the competing Spanish words need to be inhibited. It is this inhibition that has been proposed to be the reason for later retrieval difficulties for initially studied items, in this case, the Spanish words. Importantly, the need for inhibition of unwanted (Spanish) competitors in the interference phase will depend on the relative strength and ease of retrieval of items involved. This is where the frequency of use difference comes into play. Frequently used, easy to retrieve Dutch words will be less affected by competition from the recently learned Spanish words than weaker, less frequently used L2 English words. Less competition in the Dutch interference condition then requires less inhibition of the corresponding Spanish words, which in turn leads to less retrieval difficulties for these at later test, as compared to Spanish words in the English interference condition.

In Experiment 2, we attempted first of all to replicate the main effect of interference reported in Experiment 1. We also tested whether the language difference we observed in Experiment 1 can indeed be explained by frequency (of use) differences, independently of the status (native vs. non-native) of the languages involved. To do so, we manipulated word frequency *within* the participants' mother tongue Dutch. Word frequency is well known to impact ease of retrieval: low frequency words take longer to retrieve than high frequency words (Jescheniak & Levelt, 1994). We manipulated word frequency in Dutch because that allows for a maximal frequency difference between words in the low and high frequency conditions with the words still being known to the participants. Moreover, manipulating frequency within one language removes any chance for language status to play a role in driving group differences.

In Experiment 2, rather than receiving interference from different languages, all participants thus received interference in their mother tongue Dutch. However, for one group, the interferers were high frequency Dutch words (resembling the L1 interference condition in Experiment 1), while the other group received interference from low frequency Dutch words (mirroring L2 interference in Experiment 1).

If frequency of use differences are the origin of the language difference we saw in Experiment 1, we should observe a similar pattern between the low and high frequency groups as we did for the two language groups in the earlier experiment: the interference effect should be stronger for the low frequency condition than for the high frequency condition. Regardless of the frequency manipulation, we expected to replicate the main effect of interference observed in Experiment 1.

2.3 | Experiment 2

2.3.1 | Methods

The set-up of the second experiment was nearly identical to days 1 and 2 of Experiment 1. Only the differences in methods across experiments are described below.

2.3.1.1 | *Participants*

Fifty-five Dutch native speakers with normal or corrected-to-normal vision and no history of neurological or language-related disorders were recruited via the Radboud University subject pool. One participant failed to show up to the second session of the experiment, and seven (four in the high condition, three in the low condition) did not reach the learning criterion during the first session (as in Experiment 1, 30 out of 40 words), leaving 47 participants (38 female) aged 18 – 29 (M = 22.38) for analysis. All of the remaining participants reported English as their first and most frequently used foreign language in the online language background questionnaire. Proficiency self-ratings as well as performance on the English LexTALE (Lemhöfer & Broersma, 2012) are shown in Table 2.8. In contrast to Experiment 1, participants had no prior knowledge of the Spanish language, with the exception of one participant who had just started to learn Spanish via a language learning app (Duolingo); this, however, was only for two weeks. We chose participants with no knowledge of Spanish so that we could include enough high frequency words in the experiment without those words already being known to the participants. As for Experiment 1, other languages participants had learned included most prominently French, German and Latin. We stick to the terminology used in Experiment 1 and refer to Spanish as an L3 and English as the L2.

Participants were randomly assigned to one of two frequency conditions: interference from high frequency Dutch words (N = 23) or low frequency Dutch words (N = 24). The two groups did not differ in terms of English proficiency or frequency of use self-ratings, nor did they differ in their LexTALE scores or their performance on the filler tasks (all *ps* > .25).

Table 2.8

Participant characteristics – Experiment 2.

	High frequency group (N=23)			Low frequency group (N=24)		
	M	*SD*	*range*	*M*	*SD*	*Range*
English AoA	10.26	1.58	6-12	10.38	1.74	6-13
English LoE (years)	9.95	3.14	6-17	11.00	3.00	6-20
English Frequency of Use (min/day)						
Speaking	79	247	0-1200[3]	18	25	0-60
Listening	165	232	4-1200	178	162	60-720
Reading	143	194	0-900	91	91	0-300
Writing	53	102	0-480	33	57	0-240
English Proficiency[a]						
Speaking	5.13	1.17	2-7	5.04	1.08	3-7
Listening	5.69	0.82	4-7	5.71	0.95	4-7
Reading	5.95	0.64	4-7	5.87	0.85	4-7
Writing	5.13	1.06	2-6	5.00	1.06	3-7
English LexTALE	75.35	10.36	53-92	74.04	11.41	51-95
	Additional tasks					
Simon task[b]	26	14	-4-53	29	22	-16-74
Go-NoGo FAR	9	9	0-34	11	9	0-40

Note. *M* = mean; *SD* = standard deviation; AoA = Age of acquisition; LoE = length of exposure (i.e., amount of years participants had been learning English); FAR = false alarm rate (in %). [a]Proficiency self-ratings were given on a scale from 1(very poor) – 7(like a native speaker), [b]The Simon effect is expressed in ms and calculated as the difference between reaction times for the incongruent minus the congruent condition.

[3] The maximum of 1200 min is an outlier in the dataset – surely this participant did not speak English for an average of 20 hours a day. Most likely, they either typed in the wrong number by accident, or they misunderstood the question.

2.3.1.2 | *Materials*

Unlike in Experiment 1, each participant within one group (either low or high frequency) received the same set of items. Item lists thus only differed between groups with the high frequency group receiving a set of 40 high frequency (*M* = 1.50, for split by interference condition see Table 2.9) and the low frequency group receiving a set of 40 low frequency (*M* = 0.37) Dutch words chosen based on CELEX log frequencies (Dutch lemmas, Baayen et al., 1995) (see Appendix A.7 for a list of all items). Log frequencies allowed for easier matching, but see Table 2.9 for frequencies per million. The two groups of words were matched for word length in Spanish and within-group semantic similarity. Each frequency set again consisted of two subsets: 20 words that would receive interference and 20 that would not. Items in these two subsets were also matched for Spanish word length and semantic similarity (across and within sets, as in Experiment 1), and importantly also on word frequency and phonological similarity. Which set received interference was counterbalanced across participants. Finally, for the interference tasks, we also again included 20 filler items for each frequency group, which were matched for frequency to the respective target item sets and for semantic similarity (as in Experiment 1, see Appendix A.3 for details and filler characteristics). Pictures were the same as in Experiment 1; for new words, new pictures were chosen with the same selection criteria as in Experiment 1. New recordings were made for all items, again with a Spanish native speaker, this time from Andalucía (Spain).

Table 2.9
Characteristics of items used in Experiment 2.

	High frequency						Low frequency					
	Set 1			Set 2			Set 1			Set 2		
	M	*SD*	*range*	*M*	*SD*	*range*	*M*	*SD*	*range*	*M*	*SD*	*range*
Spanish word length (in syllables)	2.75	0.78	2-5	2.60	0.75	2-4	2.85	0.59	2-4	2.80	0.77	2-4
Dutch log frequency	1.61	0.26	1.23-2.26	1.75	0.50	0.95-2.95	0.33	0.32	0-1	0.27	0.32	0-1
Dutch frequency per million	50.70	45.13	17-180	115.3	197.13	9-900	2.60	2.48	0-10	2.30	2.27	0-9
Within-set semantic distance[a]	0.75	0.20	0-1.01	0.74	0.19	0-1.02	0.77	0.21	0-1.07	0.77	0.22	0-1.16
	M		*SD*		*range*		*M*		*SD*		*range*	
Spanish Levenshtein distance	5.71		1.35		2-9		6.25		1.40		2-10	
Across-set semantic distance[a]	0.82		0.12		0.49-1.17		0.79		0.10		0.44-1.03	

Note. *M* = mean, *SD* = standard deviation. Unlike in Experiment 1, participants within one group all got the same item set. Which set received interference was counterbalanced across participants. Filler items were matched in frequency and semantic similarity to the respective target item set. [a]For more information on how we controlled for semantic similarity please.

2.3.1.3 | *Procedure*

As in Experiment 1, a day after the learning session, participants returned for the interference session (High frequency group: *M* = 23.65 hours, *SD* = 2.58, range = 18-28, Low frequency group: *M* = 23.86 hours, *SD* = 2.79, range = 19-31). This time, the interference phase was in Dutch for all participants, but half the participants received interference from high frequency words, whereas the other half received interference from low frequency Dutch words. All tasks both in the learning and the interference phase were identical to Experiment 1. For both final tests, however and in contrast to the earlier experiment, we made sure that there were at most three items from the same condition in a row, and that half of the participants started the final test after interference with an interfered item, while the other half started with a not interfered item. Finally, it should be noted that there was no follow-up a week later. For feasibility reasons, and given that the interaction that we aimed to investigate further was found only on day 2 in Experiment 1, we refrained from including a day 8 session.

2.3.1.4 | *Modelling*

Responses and naming latencies were scored exactly as in Experiment 1, and were also analyzed using the same (generalized) mixed effects models as in Experiment 1, again with lme4 in R. As in Experiment 1, most errors were partial errors (83% of errors, 14% of all productions), so we again counted the number of correct and incorrect phonemes and used a two-column data frame containing these values as the input for statistical modelling. Fixed effects in this experiment were Interference (two levels: interference, no interference), Frequency (two levels: high frequency, low frequency) and their interaction. Random effects were again fitted to the maximum structure justified by the experimental design, which included random intercepts for both Subject and Item, as well as a random slope by Subject for Interference. The final random effects structure was determined based on the same principles as in Experiment 1. All p-values were calculated by model comparison (using chi-square tests).

Naming latencies again refer to the difference in naming latencies between production on the first day (after learning, serving as a baseline) and the second day.[4] As in Experiment 1, raw naming latencies on all days were first log-transformed and then difference scores were calculated and entered into the model. Fixed effects were the same as for the accuracy model and the random effects structure was also determined based on the same principles.

[4] Please see Appendix A.6 for analyses based on the raw naming latencies. The analyses on raw latencies lead to the same conclusions as those based on differences scores.

2.3.2 | Results

2.3.2.1 | *Learning Success on Day 1*

As in Experiment 1, participants were very successful at learning the new Spanish words: on average 94% (*SD* = 5%, range = 83 – 100%) of words were learned. The high and the low frequency groups did not differ in terms of learning success (Low group: 93%, High group: 94%; $t(45)$ = 0.78, p = .439, d = 0.228), or the average number of repetitions needed per item (Low group: M = 12.17, High group: M = 11.80; $t(45)$=-0.91, p = .366, d =-0.266).

2.3.2.2 | *Naming Performance After Interference (on Day 2)*

2.3.2.2.1 | Naming Accuracy

Figure 2.6 shows the mean percentages of correctly recalled phonemes in Spanish on day 2 per Interference and Frequency condition. As in Experiment 1, percentages are taken relative to the number of items learned on day 1, and are thus participant-specific. Outputs from the mixed effects model are reported in Table 2.10. We observed a main effect of Interference in line with Experiment 1: participants again made more errors on words that had received interference (7%) compared to words that had not (3%). The modulation of this interference effect by frequency was marginally significant. Separate t-tests for each frequency group showed that the interference effect was highly significant for the low frequency group (interfered: 8%, not interfered: 2%, $t(23)$ = -4.12, p < .001, d = -0.841), and marginally significant for the high frequency group (interfered: 6%, not interfered: 4%, $t(22)$ = -2.04, p = .054, d = -0.425). Though not borne out statistically in the interaction term, there is thus a trend in the predicted direction such that low frequency words tend to interfere more than high frequency words. There was no main effect of frequency.

2.3.2.2.2 | Naming Latencies

Naming latencies (in ms) at final test on day 2 per Interference and Frequency condition are plotted in Figure 2.6 and model outcomes are shown in Table 2.10. As in Experiment 1, naming latencies refer to the difference in naming speed between day 1 (baseline, after learning) and day 2, and thus reflect the slowing down of responses from baseline to final test. Again, we observed a main effect of Interference indicating that, overall, participants were slowed down more for words that had been interfered with (732 ms) than for words that had not (210 ms). There was no main effect of frequency and frequency did not significantly modulate the interference effect, although numerically there was a larger interference effect for low as compared to high frequency items.

Table 2.10

Mixed effects model outcome for naming accuracy and latencies in Experiment 2.

	Naming accuracy				Naming latencies			
Fixed effects	Estimate	*SE*	*z*	$p(\chi^2)$	Estimate	*SE*	*t*	$p(\chi^2)$
Intercept	4.31	0.27	15.82	**<.001**	0.24	0.03	7.01	**<.001**
Interference	-1.38	0.25	-5.47	**<.001**	0.25	0.03	7.67	**<.001**
Language	-0.07	0.53	-0.13	.894	0.02	0.07	0.34	.728
Interference* Frequency	-0.91	0.44	-2.04	**.057**	0.11	0.07	1.60	.106
Random effects	Groups	Var	*SD*	Corr	Groups	Var	*SD*	Corr
Item	Intercept	1.66	1.29		Intercept	0.00	0.05	
Subject	Intercept	1.88	1.37		Intercept	0.04	0.21	
	Interference	1.39	1.18	-0.38	Interference	0.02	0.51	0.26

Note. Significant effects are marked in bold. *SE* = standard error; $p(\chi^2)$ = Chi-square p-value; Var = variance; *SD* = standard deviation; Corr = correlation.

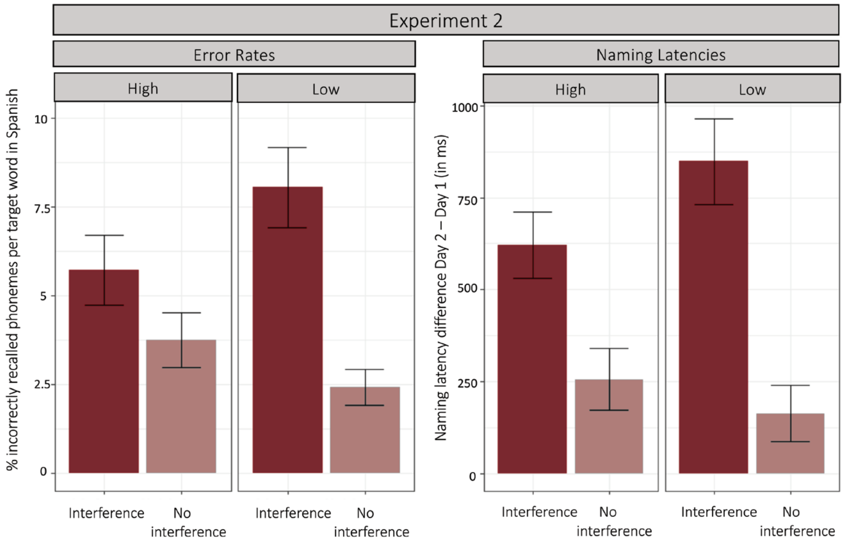

Figure 2.6

Experiment 2. Error rates and naming latencies (in ms) in Spanish productions at final test. Error rates are measured as the percentage of incorrectly recalled phonemes, and naming latencies reflect the difference in naming speed between baseline (immediately after learning) and final test. Error bars reflect the standard error around the condition means.

2.3.3 | Discussion

Experiment 2 aimed, on the one hand, at replicating the main effects of interference found in Experiment 1, and on the other hand, at understanding the language difference reported in naming latencies on day 2. Why would another foreign language (L2 English) be a stronger interferer with L3 (Spanish) word productions than the native language (Dutch)? We hypothesized that this difference was due to frequency of use differences between the languages, and that comparing interference from high vs. low frequency Dutch interferers would result in a similar pattern to that for Dutch (comparable to high-frequency) vs. English (low-frequency) interference in the earlier experiment.

We replicated the main effect of interference, both in accuracy and in naming latencies, thus lending further support to the claim that between-language interference is a driving force in FL attrition. With regard to the frequency manipulation, the results partially confirmed our expectations. At least numerically, low frequency Dutch words interfered more with L3 Spanish word productions than high frequency Dutch words. Although smaller in magnitude than the between-language manipulation, and with the current sample size only marginally significant, the frequency manipulation *within* L1 Dutch thus resulted in a pattern that resembles the *between*-language difference in Experiment 1. Given this similarity, we take the present pattern of results as partial support for the frequency of use account as a plausible explanation for the interference asymmetry in Experiment 1. Part of the reason why a foreign language interferes more than a native language with the retention of new foreign language vocabulary may thus be its relatively less frequent use, and hence that its words are harder to retrieve.

2.4 | General Discussion

In this paper we set out to study the cognitive mechanisms behind foreign language attrition. To do so, we took inspiration from the domain-general memory literature, where it has been proposed that forgetting is at least partially driven by interference from other, competing memories. In Experiment 1 we asked whether similar interference dynamics are at the basis of FL attrition and thus whether FL forgetting is driven by competition and interference from the more frequent use of other languages spoken by the individual. The results of Experiment 1 confirmed this hypothesis. Newly learnt Spanish words that had been retrieved in either L1 Dutch or L2 English were subsequently recalled less accurately and more slowly so than Spanish words that were not interfered with. These effects proved to be long-lasting, and in naming latencies on day 2, interference effects were stronger when the intermittent retrieval phase had taken place in L2 English as compared to L1 Dutch. Experiment 2 showed a comparable, albeit only marginally significant asymmetry for high- vs. low frequency words within one interference language (Dutch), suggesting that frequency of use differences might explain the differential interfering effect between native and non-native languages. Importantly, Experiment 2 also replicated the main effect of interference shown first in Experiment 1.

The main effects we report are in line with predictions made on the basis of the interference account of forgetting. Interference theory, and RIF specifically, rely on the fact that memories which share a retrieval cue (i.e., exemplars from a semantic category) compete for selection upon presentation of the shared cue. Because of this competition, repeated retrieval of one of those memories will hamper subsequent retrieval of all other, less recently retrieved items associated with the same cue (Anderson et al., 1994; Levy & Anderson, 2002). The retrieval and resulting strengthening of information can thus lead to forgetting of related memories. In the language case, these 'memories' equate to translation equivalents in the different languages a person speaks, which similarly compete with one another for selection when the speaker wants to refer to a given concept (i.e., the shared cue). The parallel activation of translation equivalents, and the possibility for resulting between-language competition, are thoroughly studied phenomena in psycholinguistics (see Kroll et al., 2008 and Chapter 1). It is very surprising that the body of literature on the possibly detrimental, long-term effects of these between-language competition dynamics is so small. Our study is among the first to show that the selective retrieval of words in one language interferes with the subsequent retrieval of the same words in other languages (as opposed to the sometimes reported translation faciliation effects, e.g., Costa et al., 1999; see Chapter 6 for a more in-depth discussion), and crucially that these effects persist well beyond the single trial and that they survive (at the least) a 20-minute delay (and thus differ from typical language switch costs; Costa & Santesteban, 2004; Meuter & Allport, 1999; and blocked switch costs; Declerck & Philipp, 2017; as well as previous language RIF studies: Levy et al., 2007, Isurin & McDonald, 2001).

We are only aware of three other studies that have attempted to draw a link between language competition and attrition. For L1 attrition, Levy et al. (2007) showed that L2 Spanish retrieval practice hampers the recall performance of L1 English words. For L2 attrition, Isurin and McDonald (2001) reported that the learning of a list of L2 Hebrew words impacts subsequent recall of a list of L2 Russian words, which was learned immediately before the Hebrew list. Finally, we recently came across a third study that looks at the effects of L1 English retrieval practice on the recall of

Welsh words learned just before, where Welsh was a previously unknown language to the participants. Bailey and Newman (2018) report longer naming latencies for interfered Welsh words as compared to not interfered words, but no effect in error rates. Results of these studies are generally in line with our results, and thus serve to reinforce the generalizability of the phenomenon.

However, there are also a number of ways in which our study differs from one or more of those three studies. First of all, as pointed out already, our study makes an important theoretical contribution by showing that the main effects of interference persist reliably (in both Experiment 1 and 2) for at least 20 minutes, a time frame that in traditional memory studies is considered to represent long-term memory, and in naming latencies even for an entire week (tested only in Experiment 1). Secondly, we allowed the newly learned L3 Spanish words to be consolidated overnight before introducing interference, which makes our design closer to real-life attrition situations. In fact, we are the first to show that interference still has an influence when the initial study (i.e., learning) and interference phases are separated by more than just a few minutes. Moreover, interference in our study comes from the mere retrieval of already known words, and not from the new learning of words, as is the case in Isurin and McDonald's (2001) study. The fact that a list learned immediately after another list overrides the first is a common finding in the memory literature (Müller & Pilzecker, 1900) referred to as retroactive interference, but seems to bear little resemblance with real-life foreign language attrition. Lastly, our study is the first to compare interference from different languages with one another, and to show that the source of interference makes a difference. Overall, we thus believe that our study provides a more realistic forgetting scenario than earlier studies on language attrition, and in doing so brings us a step closer to understanding the phenomenon and its underlying mechanisms.

2.4.1 | Does the Source of Interference Matter?

The results from Experiment 1 suggested that not only does the repeated retrieval of words in a different language hamper retrieval of L3 words, but also that it matters in which language the interference takes place. English, a foreign language to participants in our study, was a stronger interferer than L1 Dutch (in naming latencies on day 2). While this result is in line with anecdotal evidence and some previous work in psycholinguistics (Dewaele, 1998; Williams & Hammarberg, 1998), it is also at odds with the common assumption that the strong, dominant L1 interferes the most. In Experiment 2 we asked why that is the case, and how the current pattern of results could best be explained.

We hypothesized that the language differences are related to differences in their frequency of use, and resulting retrieval ease for words in these languages. Dutch, being the language of everyday interactions for our participants, is easier to retrieve than the less frequently used foreign language English. These differences in retrieval ease lead to differences in competition during the interference phase. L2 words experience more competition from the previously learnt L3 Spanish words than L1 words, thus calling for relatively more inhibition of the Spanish words in the L2 interference group. If this is true, we argued, manipulating word frequency within the mother tongue Dutch should result in a similar pattern of results. Experiment 2 suggests that this might indeed be at least part of the explanation. Low frequency words, which by our logic are comparable to L2 English words in terms of frequency and ease of access, caused at least descriptively more interference than high frequency words. Future research will be necessary to establish the reliability of this frequency by interference interaction, and indeed whether or not there is an interference effect for high frequency words (the effect within the high-frequency condition was also statistically only marginally significant).

The frequency by interference interaction, if it were to prove reliable, would resemble findings from earlier RIF studies that compared interference effects for highly prototypical category exemplars (with high taxonomic frequency, e.g., ‘orange’ and ‘apple’ for the category FRUIT) with those for exemplars of low taxonomic frequency (e.g., ‘kiwi’ and ‘papaya’). These studies report that strong exemplars suffer *more* from retrieval practice of other category exemplars than weak exemplars (Anderson et al., 1994; Hellerstedt & Johansson, 2014; though see contradictory evidence by Williams & Zacks, 2001). While this is seemingly opposite to what we report, a similar competition-based logic applies: strong representatives of a category (‘apple’ from the FRUIT category) produce more competition during the retrieval of other exemplars from their category (i.e., the interference phase) than weak representatives, and thus need to be inhibited more for successful retrieval of these other representatives. Note that the focus is here on the strength of the *competitors* during interference (which correspond to the Spanish words in our study), and not the *interferers* (the Dutch and English words in our study). However, the strengths of the two are, of course, directly dependent on each other: if an L2 English word or an exemplar of a category receives a boost through retrieval, that automatically comes at the cost of all other labels connected to the same concept (i.e., translation equivalents in L2 and L1) or exemplars connected to the same category. The magnitude of RIF ultimately depends on the *relative* strengths of competitors and interferers, differences in which can be achieved either by manipulating competitor strength, as in the RIF studies above, or interferer strength, as in our study.

Given the similarity of results in the two experiments, it seems fair to conclude tentatively that at least part of the reason why a foreign language interferes more than a native language with the retention of foreign language words is its relatively weaker, less stable status in the language system. In line with Ibrahim et al. (2017), Experiment 2 thus suggests that the differences in the strength of interference across languages observed in Experiment 1 may reflect frequency of use differences between languages. Future research should ask whether these cross-language differences can be replicated and should test further the frequency-of-use hypothesis.

2.4.2 | The Nature of Forgetting

Our measure of forgetting by interference is supported by naming accuracy on the one hand, and naming speed on the other. Naming accuracy is the most straightforward and intuitive measure of forgetting: inability to retrieve a word, or to retrieve it accurately is usually what is meant by the term ‘forgetting’ in real life. In RIF studies, accuracy in fact is usually the only measure that is reported to demonstrate interference-based forgetting. Naming latencies have only been reported in a handful of RIF studies (e.g., Bailey & Newman, 2018; Gómez-Ariza et al., 2005). Arguably, however, delayed naming latencies are a natural precursor to retrieval failure. In fact, in the psycholinguistic literature, naming latencies are the prime measure for interference effects in, for example, picture naming or language switching tasks (e.g., Costa et al., 1999, 2006; Kroll et al., 2008, 2012). Longer naming latencies in these studies are usually taken to reflect increased retrieval difficulty. Accepting that retrieval difficulty precedes retrieval inability, it follows that words that take long to retrieve might have just fallen short of being 'forgotten', and likewise that instances of retrieval failure might just reflect the extreme ends of naming latencies, indicating the point in time when an individual gives up searching for a word.

This view is very much in line with the idea that forgetting is not an ‘all or nothing’ phenomenon, but instead a gradual process described by changes in accessibility over time (see section 2.1. and also section 1.1.3 in Chapter 1). Our data further support this: while the majority of words that were forgotten on Day 2 remained forgotten on Day 8 (supporting the claim that interference can persist long term, see below), 24% of forgotten words actually recovered and were successfully retrieved on Day 8 (25% interfered, 21% not interfered). Whether this recovery reflects small, random fluctuations in activation levels for items close to the retrievability threshold (i.e., the point in time when a participant gives up searching for a word), or whether it is simply the result of re-exposure to these items during the week’s delay, is unclear. However, regardless of their origin, these conditional probabilities show that supposedly forgotten items are not necessarily lost entirely, but are in many (if not all) cases merely inaccessible and can (given favorable circumstances, such as re-exposure) be successfully retrieved again at a later point in time.

By this logic, the use of naming latencies in studies on forgetting is just as important as the use of accuracy measures, and in fact is possibly crucial to reveal subtle differences that within the context of an experimental session do not have a strong enough effect to lead to complete retrieval failure. This is furthermore especially true when participants are not given a response time limit, as in the final Spanish tests in our study. Had we set such a limit, the very long latencies, which drive the Interference by Language interaction in Experiment 1, for example, would have ended up as errors and we would likely have seen the interaction in accuracy rather than latencies (or both). In the context of our study, effects that are found only in naming latencies are thus no weaker support for our hypothesis than effects that are found (also) in accuracy (take for instance the persistence of the interference effect on day 8 only in naming latencies).

The fact that naming latency and retrieval failure (i.e., errors) are situated on a continuum also explains why we no longer observe an interaction between interference and language in naming latencies on day 8. The words that drive the language difference on day 2 are words that were correctly recalled, but that took participants a long time to retrieve (i.e., long naming latencies). If we interpret long naming latencies as the precursor to complete retrieval failure, it is these words that are the first to be forgotten between day 2 and 8. They would thus enter the analysis as forgotten words on day 8, and thus influence the accuracy statistics rather than latency statistics. In fact, our data indeed show that words known on day 2 but forgotten on day 8 took on average 1110 ms longer to retrieve on day 2 than words that were still known on day 8. By this logic, the interference by language interaction should have emerged in accuracy on day 8 instead. This was not the case. Possibly, general forgetting (in both interference conditions) washed this difference out.

2.4.3 | The Persistence of Interference

More generally, and accepting that naming latencies are just as much an indicator of forgetting as retrieval failure, our study adds to a growing body of research advocating the importance of retrieval processes in long-term memory. Next to showing that interference effects persist for at least 20 minutes, our study provides evidence for between-language competition effects that persist, for the majority of items (and thus on average), for an entire week beyond the interference induction moment, at least in naming latencies. We are aware of only a few studies that tested and showed similar truly long-term (non-language) RIF effects so far (Garcia-Bajos et al., 2009;

Storm et al., 2006, 2012).[5] The persistence of the interference effect is especially remarkable when one considers the brevity of the interference phase in our study (a mere 15 minutes of English or Dutch retrieval practice) compared to what one would encounter in real life, as well as the fact that a week of going about one's normal life introduces a lot of uncontrollable noise and, of course, natural decay of the unused Spanish words' memory traces.

Showing that language competition effects persist long term is crucial when trying to link these effects to foreign language attrition in the real world. As already discussed, in establishing between-language competition as a phenomenon with long-term ramifications, our study goes beyond language-switching studies. Besides that, our effects might also appear to resemble effects from (long-term, cross-linguistic) priming studies. For instance, Poort et al. (2016) showed that retrieval of interlingual homographs in one language leads to slower subsequent lexical decisions on the same word forms in another language. Although these inhibitory priming effects seem similar to the interference effects in our study, it is important to emphasize that they are conceptually different from the interference effects we report. Poort et al. (2016) show that it is harder to retrieve another meaning of a word with the same form (i.e., the meaning of a homograph in the unprimed language). We instead show that it is harder to retrieve another form with the same meaning (i.e., the translation equivalent of the same picture). In other words, we show inhibition on the form rather than the meaning level. Of course, it is possible that similar (or even some of the same) mechanisms that are involved in priming also underlie the effects that we report here. It remains for future research to determine to what extent that is the case. As far as language-switching is concerned, again as already discussed earlier, we believe that it is very likely that similar mechanisms are involved. What our results thus ultimately suggest is that between-language competition has both short- and long-term consequences for retrieval ease. In drawing the link between this mechanism and attrition, we hope to provide a fresh perspective to the experimental study of language attrition and to encourage future research on this topic.

[5] In order to assess whether interference induced on day 2 would persist on day 8, we tested participants twice on all of the items. This means that retrieval performance on day 8 was probably influenced by retrieval on day 2. In a future study, it might be worth increasing the number of items and testing only half of the interfered and half of the not interfered items on day 2, and the other half on day 8. Note though that there is a natural limit to how many words participants can learn within one experimental session, making such a design possibly difficult to implement.

2.4.4 | Other Directions for Future Research

From observational attrition studies we know that forgetting is often not a uniform process, but that it differs in extent from individual to individual. Though possibly less pronounced than in real life, there was also a lot of variability in individual forgetting rates in the lab-experiments reported here (Exp 1, Day 2, accuracy: -9-29%; RTs: -899-1635 ms; Day 8, accuracy: -13-27%; RTs: -1696-1687 ms; Exp 2, accuracy: -7-25%; RTs: -1595-2190 ms). It will be interesting for future studies to address these differences and to determine the factors that influence the amount to which an individual will suffer from (interference-induced) FL attrition.

An interesting candidate for an explanation of some of these individual differences is cognitive control ability. Higher cognitive/inhibitory control ability can be beneficial in that it allows for more efficient language control (Christoffels et al., 2013), but it can also have negative consequences in situations where previously irrelevant, inhibited material suddenly becomes relevant again (Treccani et al., 2009). It is possible then that participants with high inhibitory control ability suffer *more* from language-RIF because they more efficiently inhibit Spanish competitors during English/Dutch retrieval in the interference phase, making the Spanish words subsequently harder to recall. Exploratory analyses using Simon and Go-NoGo task performance as predictors in the statistical models for both experiments, however, did not lend consistent support to this hypothesis (see Appendix A.8 for the results). Our experiments were not designed to accommodate individual difference analyses though, neither in terms of sample size, nor in terms of experimental set-up (both tasks were included as filler tasks and used to match participants across groups). We thus leave it to future studies to test for effects of cognitive control ability. Should this ability prove relevant for induced attrition in the lab, it would also be interesting to include it in the standard test battery in observational attrition studies.

Along similar lines, it might be worth asking whether the amount of previously learned foreign languages, and the level of proficiency reached in those languages, impacts an individual's susceptibility to interference. People who have ample experience in multiple foreign languages might be more experienced at dealing with language interference and hence less prone to suffer from it. Our experiments were again not designed to answer this question, especially because there was very little variability in our sample with regard to the number of already known foreign languages (M = 2.77, SD = 0.87, range = 1-4, also see Appendix A.8 for histograms for each experiment). Future research should sample participants accordingly to disentangle the role of degree of multilingualism in FL attrition.

Relatedly, age of onset (AoA) of bilingualism might play a role. Costa and Santesteban (2004) argued that late bilinguals rely more on inhibitory control of non-target languages in speech production than early and highly proficient bilinguals. Such a difference in reliance on inhibitory control as a mechanism to switch between languages might again translate to differences in interference susceptibility. In exploratory analyses for Experiment 2 (though not for Experiment 1), we indeed found that participants who started learning foreign languages earlier on in life showed smaller interference effects than late bilinguals (see Appendix A.8). Just as the other individual difference analyses mentioned above, this result should, however, be taken with a grain of salt, especially given that it is not consistent across experiments. Future research will need to replicate this finding before conclusions can be drawn based on the direction of the effect.

Moving away from individual differences, there are other aspects of the design that could be adjusted in future studies, which might further help understand and disentangle the nature of the interactions in our experiments. Frequency of use and language status (native vs. non-native) are confounded with age of acquisition in our experiment: all of our participants live in their L1 environment and so their L1 is both their first acquired language *and* the most frequently used language in daily life. It would be interesting to repeat Experiment 1 with participants who are immersed in an L2 environment, for whom the L1 would still be the first acquired language, but no longer the most frequently encountered language in daily life. If, as Experiment 2 suggests, frequency of use is the main determinant of interference strength for a given language, the pattern of results should reverse in an L2-immersion setting. Such a finding would further strengthen the claim made by Ibrahim et al. (2017) that processing asymmetries between native and non-native languages can be traced back to frequency of use differences, as well as of course the conclusions we draw on the basis of Experiment 2 in the present chapter.

2.4.5 | Conclusions

The experiments reported in this chapter show that foreign language attrition is (at least partially) caused by retrieval competition dynamics between languages. More specifically, the retrieval and practice of translation equivalents from other languages interferes with the future retrieval of words in the target foreign language. Such interference effects are strongest between foreign languages. Finally, we show that between-language interference effects are not just momentary forgetting effects, but in fact are long-lasting, and thus make for a plausible mechanism to account for foreign language attrition as it occurs in the wild.

CHAPTER 3

Electrophysiological Evidence for Cross-Language Interference in Foreign-Language Attrition

This chapter is based on:

Mickan, A., McQueen, J. M., Valentini, B., Piai, V., & Lemhöfer, K. (2020). *Electrophysiological evidence for cross-language interference in foreign-language attrition.* Manuscript submitted for publication.

I would like to thank Beatrice Valentini and Giulio Severijnen for help with data acquisition and preprocessing as well as Rosarita Leone for the Italian audio recordings and Vitória Piai for advice on data analysis.

Abstract

Foreign language attrition (FLA) appears to be driven by interference from other, more recently-used languages (Chapter 2). Here we tracked these interference dynamics electrophysiologically to further our understanding of the underlying processes. Twenty-seven Dutch native speakers learned 70 new Italian words over two days. On a third day, EEG was recorded as they performed naming tasks on half of these words in English and, finally, as their memory for all the Italian words was tested in a picture-naming task. Replicating Chapter 2, recall was slower and tended to be less complete for Italian words that were interfered with (i.e., named in English) than for words that were not. These behavioral interference effects were accompanied by an enhanced frontal N2 and a decreased late positivity (LPC) for interfered compared to not interfered items. Moreover, interfered items elicited more theta power. We also found an increased N2 during the interference phase for items that participants were later slower to retrieve in Italian. We interpret the N2 and theta effects as markers of interference, in line with the idea that Italian retrieval at final test is hampered by competition from recently practiced English translations. The LPC, in turn, reflects the consequences of interference: the reduced accessibility of interfered Italian labels. Finally, that retrieval ease at final test was related to the degree of interference during previous English retrieval shows that FLA is already set in motion during the interference phase, and hence can be the direct consequence of using other languages.

3.1 | Introduction

Most people who have learned a foreign language will be familiar with the frustrating feeling of losing access to that language over time, no matter how much effort they put into learning it in the first place. Why this happens, and why the so-called attrition process appears to be so inevitable, is a long-standing issue in the language sciences. Recent research suggests that foreign language attrition can be the direct consequence of using and speaking other languages (e.g., Chapter 2, Levy et al., 2007). In Chapter 2, for example, we showed that the mere act of retrieving words in either a native or a foreign language hampers subsequent access to translation equivalents in another foreign language, and that these interference effects are long-lasting. The neural correlates of these processes and, hence, their exact contribution to foreign language attrition, however, are still unknown. The current study aims at filling this gap. Building on Chapter 2, we seek to establish the electrophysiological correlates of between-language interference. The electroencephalogram (EEG) provides a different way of looking at the attrition process, and, crucially, allows us to understand precisely when in time these interference effects emerge. In doing so, we hope to shed light on the temporal dynamics of the underlying mechanisms of interference-based foreign language (FL) forgetting.

3.1.1 | Between-Language Competition and Inhibition as Driving Forces in FL Attrition

Participants in Chapter 2 first learned a set of new L3 Spanish words. One day later, they were asked to repeatedly retrieve half of those in either L1 Dutch or L2 English. Finally, after a delay of 20 minutes, participants were tested again on all originally learned Spanish words. Naming performance in this final test showed that people were significantly worse at recalling words that they had just named in English or Dutch: they made more mistakes and were slower to recall interfered compared to not interfered items in Spanish. Lexical retrieval of translation equivalents in a different language can thus make you forget the same words from a (recently learned) foreign language. In reaction times, this effect persisted until a week after interference induction, thus providing the first evidence for true long-term effects of between-language interference, and hence establishing it as a plausible mechanism of foreign language attrition.

These and comparable retrieval-induced forgetting effects in the memory literature tend to be explained through competition and inhibition processes (Anderson et al., 1994). In the language case, specifically, Chapter 2 built on the assumption that translation equivalents in different languages compete for access during lexical selection and that words in competing, non-target languages are inhibited to allow for efficient communication in the target language (see Kroll et al., 2008). In explaining our findings in Chapter 2, we thus reasoned that during the retrieval of English words in the interference phase, for example, the recently learned Spanish words competed for access with their English translation equivalents and hence that they had to be inhibited for successful retrieval of the latter. Assuming that this inhibition is long-lasting, it should result in a competition disadvantage for the suppressed Spanish items at delayed final test (i.e., after 20 minutes). In order to retrieve the Spanish labels, their inhibition first needs to be lifted and competition from their recently retrieved English competitors needs to be overcome, which takes time and hence slows down or entirely blocks retrieval. Words in the no-interference condition, which were not retrieved in English and hence did not need to be inhibited in Spanish, should consequently be easier to retrieve and experience less interference from English competitors at final test than items that were interfered with, which explains why the former were recalled faster and more accurately than the latter in Chapter 2.

In the present chapter, we sought to replicate the main effects of interference reported in Chapter 2, but crucially aimed to further our understanding of the underlying cognitive mechanisms driving this behavioral effect. To that end, we measured EEG activity both during the interference phase and during the final test phase and asked whether we could track the competition and inhibition dynamics that are often called upon in explaining the behavioral between-language interference effects. Instead of testing for interference on L3 Spanish, we used Italian as L3 in the current study, and the interference phase consisted only of L2 English retrieval practice (leaving out the L1 Dutch group from Chapter 2). Dutch native speakers thus first learned a set of words in L3 Italian, and subsequently, a day later, retrieved half of them in L2 English, before being tested again on all originally learned Italian words.

We expected to observe neural correlates of interference and inhibition both at final test and during the interference phase. Moreover, in looking not only at the outcome of such interference (i.e., the final test), but also at the moment in time when forgetting is supposedly induced (i.e., the interference phase), we aimed at testing the assumption that activity during the earlier phase is directly related to performance at final test. If competition and inhibition during the interference phase indeed predict retrieval ability at final test, we should be able to observe more competition/inhibition for items that are later on slower to retrieve (i.e., harder to recall at final test) compared to items that are fast to retrieve at final test.

3.1.2 | Stimulus-Locked Neural Markers of Interference, Competition and Inhibition

3.1.2.1 | *Evidence from Event-Related Potentials*

In event-related potentials (ERPs) in the EEG, inhibition and interference are commonly associated with an early anterior negative deflection, the so-called N2 component. Maximal over frontal electrode sites and peaking between 200 and 350 ms (time-locked to stimulus presentation), this component has frequently been observed in studies using the language switching paradigm, where people alternate between naming pictures in their L1 and L2 prompted by a language cue. In those situations, it is common to find an enhanced N2 for switch trials, where the language of naming differs from the previous trial, compared to repeat trials where the language remains the same (Jackson et al., 2001; Zheng et al., 2020; but see Christoffels et al., 2007, for a larger N2 for repeat compared to switch trials). These N2 switch costs are typically interpreted to reflect inhibition of the non-target language, in line with interpretations of comparable N2 findings in non-linguistic tasks that require inhibition of a prepotent response (e.g., the Go-NoGo-task; see Folstein & Van Petten, 2008, for a review). Some researchers have instead argued that the N2 is a signature of response conflict, rather than evidence for the resolution of that conflict (i.e., through inhibition of interfering responses or boosting of target responses; Nieuwenhuis et al., 2003). Crucially though, both interpretations assume that it is an indicator for the presence of interference, and hence is a viable candidate for a neural correlate of interference-based foreign language attrition.

It should be noted that most evidence for language switch N2 effects comes from mixed-language switching paradigms, which differ in design from the current study in important ways. First of all, traditional language switching studies test for inhibition on a global, whole-language level rather than locally on the item level: they ask what naming a picture in, for example, L1 does to subsequent naming of any other picture in L2, rather than to naming of its L2 translation equivalent. Moreover, they observe the effects of language switching from one trial to the next, but not their potential long-term effects (though see Branzi et al., 2014; Misra et al., 2012; Wodniecka et al., 2020; reviewed in detail in the section 3.4). It remains to be seen whether the sustained, local interference / inhibition effects underlying foreign language attrition are reflected in the same N2 modulation as the short-lived, global effects in mixed-language switching studies. Finally, language switching studies differ from our study in that they target switching between two already known languages, namely L1 and L2, but not, at least to our knowledge, switching between two foreign languages, of which one has just recently been learned. The neural correlates of the consequences of naming in one foreign language on subsequent, delayed naming in another, just recently learned foreign language, as studied here, thus remain to be investigated. If the effects observed in Chapter 2 are caused by language interference and inhibition and assuming that the N2 reflects these processes not just globally, but also locally, we should expect modulations of the N2 component in the EEG during both the interference and the final test phase.

Another component that is sometimes reported in language switching studies is the LPC, a late positive component with a posterior parietal topography, occurring between 300 and 900 ms post stimulus onset. Just like the N2, the LPC is bigger for switch compared to non-switch trials and has hence been interpreted as a continuation thereof, indexing the after-effects of language interference and inhibition (Jackson et al., 2001). This component is not always found, and in fact not even always inspected (the time window for analysis in switching studies is typically limited to the first 500 ms post stimulus presentation), and hence it is unclear how robust this signature is. Importantly, this ‘switching LPC’ is not to be confused with the much more frequently reported LPC in the memory literature. The ‘memory LPC’ is thought to reflect long-term episodic recognition processes. During retrieval, it has been reliably found to be bigger for old compared to new items (for a review, see Rugg & Curran, 2007). In contrast to the switching LPC, which appears to be enhanced during retrieval of items on which there is more interference (i.e., from a non-target language, namely on switch trials), the memory LPC is found to be stronger for items where retrieval is more accurate and successful (e.g., Finnigan et al., 2002; Rugg et al., 1995; Wilding, 2000), and hence would be expected to show the opposite pattern, that is to be enhanced in trials where interference is low rather than high. Note, however, that this memory LPC is typically reported in recognition paradigms, rather than during productive recall. It remains to be seen whether our interference manipulation influences either of these late positive effects and, hence, whether the LPC is also a marker of foreign language attrition or not.

3.1.2.2 | *Evidence from Neuronal Oscillations*

In the frequency domain, interference has been consistently associated with power increases in the theta band (4-7 Hz) of the EEG signal. Evidence comes, for example, from studies using tasks with response conflicts, such as the Go-NoGo or Stroop tasks, where theta power (time-locked to stimulus onset) is enhanced in the conflicting compared to the not (or less) conflicting condition (Hanslmayr et al., 2008; Nigbur et al., 2011). These theta effects occur anywhere within the first 1000 ms post stimulus presentation, tend to have a mid-frontal scalp distribution and are thought to reflect interference from alternative responses, and possibly the recruitment of executive control processes to overcome this interference.[6]

In the language domain, theta power has been linked to semantic interference. Naming a picture with a semantically related, same-language distractor displayed on top triggered more theta activity than naming a picture with a semantically unrelated, and hence less interfering distractor on top (Piai et al., 2014). Between-language interference, time-locked to the presentation of a stimulus, as targeted in this paper, however, has not yet been linked to theta power increases. To our knowledge, there are no studies on the oscillatory dynamics of stimulus-induced between-language competition in bilingual word production.

Further evidence for theta as a marker for interference magnitude comes from memory studies on retrieval-induced forgetting (RIF; e.g., Ferreira et al., 2014; Hanslmayr et al., 2010; Staudigl et al., 2010). These studies typically contrast competitive and non-competitive interference conditions during the retrieval of previously learned category-exemplar associations. Staudigl et al. (2010), for example, had participants either actively retrieve a subset of previously studied exemplars from a given category, or passively restudy category-exemplar pairs. In the active retrieval condition, the presentation of the category cue activates other exemplars which compete with selection of the to-be-retrieved exemplar, while no such competition and interference emerges when passively viewing category-exemplar pairs. In line with the idea that theta is a marker for interference magnitude, theta power was found to be increased during retrieval in the active retrieval task as compared to the passive exposure task. Changes in theta power from the first to the second round of active competitive retrieval were furthermore found to be related to later forgetting. Forgetting in RIF studies is measured in a final test on all originally learned category-exemplar pairs, both those intermittently retrieved or restudied and those not part of the interference phase at all. Behaviorally, Staudigl et al. (2010) only observed forgetting for exemplars whose competitors (i.e., other exemplars from the same category) were actively retrieved in the interference phase, but not for exemplars whose competitors were only restudied. Crucially, the magnitude of forgetting was positively correlated with the decrease in theta from the first to the second round of retrieval practice, suggesting that interfering competitors were suppressed during competitive retrieval and that the amount of this suppression was related to later forgetting.

[6] Note that these theta effects are different from the theta effects that have been linked to successful memory encoding; these will not be discussed here any further (e.g., Klimesch et al., 1996).

Next to oscillations, EEG RIF studies also sometimes report ERPs (Ferreira et al., 2014; Hanslmayr et al., 2010; Johansson et al., 2007). Unlike the theta effects, the ERP signatures they report vary considerably from study to study though, ranging from prolonged positivities (Johansson et al., 2007) to a combination of short-lived posterior negativities and anterior positivities (Ferreira et al., 2014; Hanslmayr et al., 2010) for competitive compared to non-competitive retrieval. Overall, it should be noted that comparisons in EEG RIF studies are often between entirely different tasks (e.g., active retrieval vs. passive restudy), making it unclear to what extent their theta and ERP signatures reflect only competition or also other task-related differences between conditions. Even when the comparison is between two active retrieval tasks though, as in Hanslmayr et al. (2010), their stimuli (category-exemplar pairs) and task design (covert rather than overt retrieval) make the comparison to the present study difficult. Given these design differences and the inconsistent ERP signatures RIF studies report, it is questionable how relevant they are for hypothesis formulation for the present study. For ERPs, we consider the N2 component to be much more likely given its reliable presence in studies that require switching between languages in overt picture naming paradigms. For theta oscillations, there is no evidence for their involvement in competitive bilingual lexical retrieval to this point, and hence it will be interesting to see whether they are implicated in the type of between-language competition and interference that is supposedly at play in foreign language attrition, or not.

3.1.3 | The Present Study

To sum up, the present study investigates the neural correlates of foreign language attrition. Building on previous behavioral studies, we sought converging neural evidence for between-language interference and inhibition as driving forces behind foreign language vocabulary forgetting. To that end, Dutch native speakers first learned 70 new Italian words over the course of two consecutive days. On a third and last day, they were asked to retrieve half of the learned words in English, a foreign language they already knew, and were subsequently tested on all originally learned Italian words. We chose English as interference language because in Chapter 2 we had found that foreign languages tend to be stronger interferers than the L1. We measured EEG during all sessions on the third day, that is both during the picture naming tasks in the interference phase and at final test.

Behaviorally, we expected to replicate Chapter 2 despite the change in language (Italian rather than Spanish), the extended memory set (70 instead of 40 to be learned words) and the fact that the learning session was spread over two rather than just one day (to compensate for the bigger memory set). We thus predicted to observe more errors and slower naming responses to interfered than not interfered words at final test in Italian. Critically, though, based on the EEG literature reviewed in the previous sections, we also expected those behavioral effects to be accompanied by possibly more theta power and most likely an increased N2 component for interfered items at final test. In line with how these signatures are typically interpreted, we hypothesize that theta indexes the interference that the Italian items experience from the recent practice of their English translation equivalents and that the N2 reflects the higher need for inhibition of the latter to resolve this interference. We had no clear expectations with regard to the LPC.

Finally, we were also interested in the interference phase itself, when forgetting is supposedly induced. Here our comparison of interest concerned only the items in the interference condition (as the other items did not occur in this phase). If cognitive control dynamics during the interference phase are responsible for performance deficits at final test, we should observe more evidence for such processes on items that are later more difficult to recall at final test. To that end, we analyzed, per participant, their trials in the interference phase based on a median split of their naming latencies at final test. We expected that words that took participants longer to recall at final test would show an enhanced N2 and stronger oscillations in the theta range during the English interference phase than words that participants were faster to retrieve at final test in Italian.

3.2 | Methods

3.2.1 | Participants

Thirty Dutch native speakers were recruited via the Radboud University participant pool. One failed to reach the learning criterion on day 2 (see section 3.2.3 for details), and hence had to be excluded from the remainder of the study. Two additional participants had to be excluded from analysis because they had too many EEG artifacts (see section 3.3.1 for details). The remaining 27 participants (18 female) were between 18 and 26 years old (*M* = 21.07; *SD* = 2.00). All of them were right-handed, had normal or corrected-to-normal vision, and reported no history of language-related or neurological impairments. For the analysis of the interference phase EEG recordings, one of these 27 participants had to be discarded because of technical failure (and hence missing data) during this part of the experiment.

Before coming to the lab, participants were asked to fill in an online language background questionnaire. This was done to ensure that our participants had no (or minimal) prior knowledge of Italian. Only one participant reported prior knowledge of Italian. He had only just started learning Italian on Duolingo a month prior to participating in the study, and judged his Italian as very poor (1 out of 7). He was deemed sufficiently inexperienced with Italian to still be included in the experiment.

As also established through this online questionnaire, Dutch was our participants' only mother tongue and English was the first learned foreign language for all participants. Table 3.1 summarizes our participants' frequency of use and proficiency self-ratings in English, as well as their performance on the English LexTALE, a standardized lexical-decision based vocabulary test (Lemhöfer & Broersma, 2012). Other languages participants spoke included most prominently German, French and Latin. We refer to Italian as an L3 because it was learned after L2 English. For some participants, Italian was actually L4 or even L5, but we stick to L3 for simplicity.

Table 3.1

Participant characteristics.

	M	*SD*	*range*
English AoA	10.41	1.19	8-12
English LoE (years)	9.67	3.09	5-16
English Frequency of Use (min/day)			
Speaking	29	62	0-300
Listening	160	132	0-480
Reading	87	74	0-270
Writing	43	67	0-240
English Proficiency[a]			
Speaking	5.52	1.09	4-7
Listening	5.93	1.00	4-7
Reading	5.97	0.90	4-7
Writing	5.37	1.25	3-7
English LexTALE	74	14	43-95

Note. *M* = mean; *SD* = standard deviation; AoA = Age of acquisition; LoE = length of exposure (i.e., amount of years participants had been learning English). [a]Proficiency self-ratings were given on a scale from 1 (very poor) to 7 (like a native speaker).

Participants gave informed consent and received either course credit or vouchers for their participation (10€/h). The study was approved by the Ethics Committee of the Faculty of Social Sciences, Radboud University.

3.2.2 | Materials

Participants learned 70 Italian nouns referring to concrete, everyday objects or animals (see Appendix B.1 for the list of words). All words were non-cognates between Italian, Dutch and English, and were between two and four syllables long in Italian (*M* = 2.69, *SD* = 0.50) and between one and three syllables long in English (*M* = 1.33, *SD* = 0.67). Their corresponding Dutch lemma frequencies ranged from 0 to 180 per million (*M* = 25.11, *SD* = 37.30; CELEX, Baayen et al., 1995). Pictures for each of the words were chosen from Google (www.google.com) and the BOSS database (Brodeur et al., 2010). They were photographs of the respective object or animal centered on a white background (6x6 cm) and adjusted for size so that they occupied a maximum of 400 px in either width or length. Finally, each noun was recorded by a female Italian native speaker from Rome (Italy).

These 70 words were subdivided into two subsets of 35 words each: one of those two subsets was later interfered with, that is retrieved in English on day 3, while the other was not (see Appendix B.1 for each word's set assignment). Which set received interference was counterbalanced across participants. Importantly, though, the two subsets were matched in terms of word length in both Italian and English, Dutch frequency, within-set phonological similarity as assessed via Levenshtein distances (Levenshtein, 1966), and within-set semantic similarity (expressed as a distance value derived from semantic vectors with smaller values corresponding to high semantic similarity, as described in Mandera et al., 2017) (see Table 3.2 for averages of these values per set).

For the interference phase, 35 filler items to be named in English were chosen in addition to the 35 experimental items that would receive interference. Filler items were not analyzed, and were merely included to disguise the fact that only half of the originally learned experimental items were part of the interference session. Filler items were nevertheless matched to the experimental items in terms of English word length (*M* = 1.43, *SD* = 0.50, range = 1-2) and Dutch frequency (*M* = 1.20, *SD* = 0.55, range = 0-2.24).

Table 3.2

Item characteristics.

	Set 1			Set 2		
	M	*SD*	*range*	*M*	*SD*	*range*
Italian word length (in syllables)	2.66	0.68	2-4	2.72	0.66	2-4
English word length (in syllables)	1.26	0.44	1-2	1.40	0.55	1-3
Dutch Celex log frequency	0.97	0.66	0-2.14	1.09	0.54	0-2.26
Dutch Celex per million frequency	24.63	34.36	0-137	25.60	40.53	0-180
Semantic distance[a]	0.81	0.17	0-1.09	0.81	0.17	0-1.05
Italian Levenshtein distance	6.42	1.64	2-11	6.28	1.51	1-10

Note. *M* = mean, *SD* = standard deviation. Which set received interference was counterbalanced across participants. [a]Semantic similarity was assessed via semantic vectors, as described in Mandera et al. (2017). Small values reflect higher semantic overlap.

3.2.3 | Procedure

The study consisted of three consecutive testing days (see Figure 3.1 for a schematic overview). On the first two of those, participants were asked to learn 70 Italian words via a mix of receptive and productive tasks with feedback. Learning success was established at the end of each day via a post-test without feedback (see below for details). No EEG was acquired during either of those two days. The third and last day started with the so-called interference phase, during which participants were asked to retrieve half of the learned words in English, and ended with a final test in Italian on all originally learned words. To avoid confusion, we refer to the Italian recall test on day 3 as 'final test', rather than post-test; the word 'post-test' is used to refer to the Italian recall tests at the end of days 1 and 2. EEG was acquired during the entire session on day 3, that is both during the interference phase and the final test in Italian. Below, we will describe the tasks of the various phases of the experiment in more detail.

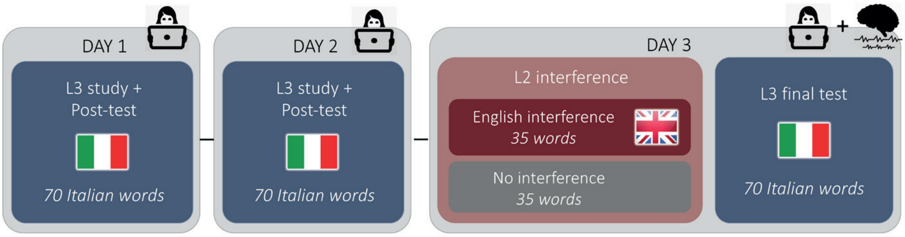

Figure 3.1

Schematic overview over experimental set-up.

All tasks were administered using Presentation (Version 19.0, Neurobehavioral Systems, Inc., Berkeley, CA) on a Dell T3610 computer (3,7Ghz Intel Quad Core, 8GB RAM, Windows 7, monitor: BenQ XL 2420Z, 24-in, 1920 x 1080 pixels, 120 Hz refresh rate). All audio stimuli were presented to the participants via headphones (Sennheiser HD201), and all oral responses were recorded via a microphone (Shure 16a) in WAV format using a Behringer X-Air XR18 digital mixer.

On all days, participants were tested individually in a quiet room. For the behavioral sessions on days 1 and 2, the experimenter sat in a room next to the participant's room. The door between these two rooms was kept open at all times for efficient communication, and for the experimenter to be able to code the participant's responses. On day 3, for the EEG session, the experimenter also sat in an adjacent room, this time the door was kept shut during the experiment and communication between experimenter and participant was done via microphone.

3.2.3.1 | *Day 1 – Italian Learning Phase 1*

As in Chapter 2, the learning phase consisted of a series of receptive and productive tasks that started out easy, and got progressively more engaging and difficult. The tasks in the learning phase were identical in terms of stimulus timings and set up to the learning tasks used in Chapter 2 (for procedural details additional to those reported below, please see Appendix A.4).

The learning phase on day 1 started with a familiarization round, during which participants listened to and saw (written versions of) each of the 70 words on screen, as well as their corresponding pictures once. Participants clicked through the pictures at their own pace. Next to acquainting themselves with the items, they were also instructed to let the experimenter know if they already knew any of the Italian words. Italian words that were already known to a participant were later excluded from analysis (also see section 3.3.1; number of excluded items across participants: *M* = 0.67, *SD* = 1.71, range = 0-8). This initial familiarization round was followed by two rounds of a two-alternative forced choice task, in which participants saw all 70 pictures twice, each time with two Italian labels from the list of to-be-learned words underneath. Participants were asked to choose the word that matched the picture they saw by clicking on it with a mouse. They then received automatic feedback on their performance (a green or red square around the picture for correct or incorrect responses respectively, accompanied by the correct word underneath the picture and its corresponding audio). After the feedback, the next trial started automatically. In the second round of this task, before seeing the two labels, participants were asked to guess the Italian word for the picture and say it out loud. This was done to start engaging them more actively with the words. The experimenter initiated the appearance of the two labels after a participant had made an attempt at naming the picture, and the rest of the trial continued as in the first round.

Next, participants completed two rounds of a word completion task. They saw each of the pictures together with their respective first syllables (or first grapheme for monosyllabic words) and were asked to complete the word out loud. The experimenter coded their answers for correctness (either as fully correct, fully incorrect or partially correct), which initiated feedback (identical to the feedback in the multiple-choice task, a green frame was displayed only for fully correct answers). From this task onwards, participants could decide for themselves when to continue with the next trial; they were thus allowed as much time as they needed to process the feedback. The word completion task was followed by a writing task. For this task, participants saw each picture once and were asked to write down (on a piece of paper) the Italian word for the picture. They then hit the enter key, which initiated the visual presentation of the correct Italian label and its spoken form. Based on this feedback, participants then corrected themselves, when necessary, by writing down the correct word on the same piece of paper. They were instructed to write each word on a new piece of paper, and to turn over each piece after use so that they would not be able to see their earlier responses.

The first day ended with two rounds of picture naming. The first of those rounds was still with feedback: participants saw each picture once, had to name it and received feedback initiated by the experimenter. Words were again coded as fully correct, fully incorrect or partially incorrect, and feedback was also again a green (for fully correct answers) or red screen (for partially and fully incorrect answers) together with the correct label (presented visually and auditorily). The second round served as a post-test, to establish which words had already been learned. During this last round, participants no longer received feedback.

3.2.3.2 | *Day 2 – Italian Learning Phase 2*

Day 2 (mean hours between day 1 and day 2 = 23.17, *SD* = 2.78, range =18-29) of the learning phase started with another round of the word completion task. The set-up was identical to the word completion task on day 1. The remainder of the session was spent with picture naming tasks, similar in set-up to the picture naming tasks on day 1. Participants named each picture at least twice with feedback. If a participant knew all words in those two rounds, they proceeded to one more round of naming with feedback, followed by one final round of naming without feedback (i.e., the second Italian post-test, identical in set-up to the post-test on day 1). If a participant did not know all words during the first two rounds of picture naming, the unknown words were repeated until he/she had named all pictures correctly in at least two consecutive rounds. Each repetition round consisted of at least ten pictures: if a participant only had two pictures left to learn, they would thus get both these pictures, but also eight already known pictures to name. This was done to ensure sufficient difficulty in naming even when there were only few items left to be learned. Repetitions of already known words were counterbalanced, such that each word was repeated approximately equally many times.

Throughout the entire learning session, taking day 1 and 2 together, participants saw each picture minimally 14 times (*M* = 15.34, *SD* = 1.03, abs. range = 14-30). Both sessions took a maximum of 1.5 hours. If on day 2 participants were still on the adaptive picture naming task after 1 hour and 15 minutes, this task was stopped by the experimenter, and the remaining two rounds of picture naming were administered. Participants were required to learn a minimum of 50 out of the 70 words (spread equally over conditions) to be able to continue to day 3. As mentioned in section 3.2.1, all but one participant reached this criterion.

The order of items in all learning tasks was participant- and task-specific. To avoid order effects during learning, pictures were presented in a different, random order in each task. To keep the distance between item repetitions constant within a task though, the order of items was identical for consecutive rounds of a single task. Due to the set-up of the tasks, there were never more than two identical rounds in a row. For the two post-tests (one on each day), six lists were created, making sure that no more than three items from the same condition (interfered / not interfered) followed in immediate succession, and that half of the participants started each post-test with an item from the interference condition and the other half with an item from the no interference condition. The items in lists 4 to 6 followed the reversed order of the items in lists 1 to 3. Each participant got a different list for each of the two post-tests, but never two reversed lists (e.g., never 1 and 4, or 2 and 5).

3.2.3.3 | *Day 3 – English Interference Phase and Final Italian Test*

3.2.3.3.1 | Interference Phase

Day 3 (mean hours between day 2 and day 3 = 24.46; *SD* = 3.62, range = 17-32) started with an interference phase, during which participants saw the pictures corresponding to half of the learned Italian words, as well as 35 filler pictures, and had to retrieve the names of the pictures in English. In total, they saw each picture nine times: once during an initial (English) familiarization task with feedback, four times during a picture naming task without feedback and another four times during a letter search task without feedback. EEG data were acquired during all these tasks but only those from the picture naming tasks were analyzed. Furthermore, out of the picture naming tasks, only the first and last rounds were analyzed (see section 3.3.3.2 for details).

In the familiarization task, each trial started with a fixation cross presented on the screen for 1500 ms, followed by a blank screen for 500 ms, followed in turn by the picture together with the first syllable (or the first grapheme in case of monosyllabic words) of its corresponding English label. We chose to present syllables rather than the initial letter to make naming easier. The picture and text were displayed for 2000 ms. Participants were instructed to withhold their response during this delay period. In the subsequent picture naming tasks, this delay would serve as the time window for EEG analysis and hence needed to be as free of movement artifacts as possible. Given that this delay is the same for all words regardless of which condition they belong to, differences between conditions should be unaffected by it. Data of the familiarization task were not analyzed, and hence the delay was not strictly necessary here, but we included it anyway to familiarize participants with the task timing. After these 2000 ms, a question mark appeared on the screen prompting the participant to give their response, that is, name the picture in English. The experimenter coded their answers for correctness (fully correct, fully incorrect or partially correct as during the learning phase), and in doing so initiated a feedback screen, which unlike the feedback in the learning tasks only contained the intended, correct English label for the picture, but no green or red screen and also no audio. If a participant had been unable to name a picture in English, the experimenter asked whether they at least recognized the word on screen, or whether it was indeed an entirely unknown English word for the participant. If a participant indicated recognizing the picture, the item was subsequently marked as known rather than unknown. Only truly unknown words, that is target words that were neither named correctly nor recognized by a participant, were later excluded from analysis in all tasks (see section 3.3.1; average number of excluded items: $M = 1.33$, $SD = 2.10$, range = 0-9). The feedback screen remained visible until the experimenter confirmed or changed the correctness coding. The next trial then started automatically.

The picture naming task also started with a 1500 ms fixation cross, followed by a 500 ms blank screen, and finally the picture for 2000 ms. Participants were again instructed to withhold their response during this delay window, and to blink as little as possible. The experimenter again coded responses for correctness, but participants did not receive feedback, and the experimenter's button press immediately initiated the next trial.

In the letter search task that followed, participants had to decide whether or not the English word for the picture contained a certain letter. For each round, participants got a new letter (one of R, L, T, or N). A trial started with a 500 ms fixation cross, followed by a 250 ms blank screen, and finally the picture, which remained on screen until a participant pressed a button (right button for yes, left button for no), or for a maximum of ten seconds. Participants did not receive feedback on their performance.

In order to make the interference phase less monotonous, we split the picture naming and letter search tasks, such that participants first underwent two rounds of picture naming, followed by two rounds of letter searching (letters R and L), followed by two more rounds of picture naming and two more rounds of letter searching (letters T and N). The presentation order of items in the interference tasks was semi-randomized. For the familiarization task, each participant was assigned to one of eight lists, making sure that no more than three items from the same condition (filler vs. experimental items) appeared in immediate succession, and that half of the participants started the task with a filler word and the other half with a target item from the interference condition. For the picture naming task, the same restrictions held. Here participants got two of eight lists, one for each block (one block consisting of two rounds), ensuring that they did not get the same list in the two blocks. For the letter search task, the order of items was semi-random, ensuring that no more than three 'yes' or 'no' responses followed in immediate succession.

3.2.3.3.2 | **Filler Task – Go NoGo**

To temporally separate the interference phase from the final test, following Chapter 2, participants completed a 20-minute long Go-NoGo task after interference and before the final test in Italian (based on Nigbur et al., 2011, the only difference being that stimuli remained on screen for a maximum of 1000 ms rather than just 200 ms). No-Go false alarm rate was on average 4% (*SD* = 5%, range = 0-24%). Since this task merely served as a filler task, we did not analyze the data any further.

3.2.3.3.3 | **Final Italian Test**

Finally, to assess what interference did to participants' Italian knowledge, participants were tested again in Italian on all 70 originally learned words. Participants were asked to name all pictures twice. We chose for two rounds of naming because of possible recency of exposure differences between interfered and not interfered pictures, which the EEG is sensitive to. In ERPs, (recently) repeated words and pictures (e.g., faces) elicit attenuated N400s and enhanced LPCs compared to nonrepeated words and pictures (e.g., Bentin & McCarthy, 1994; Rugg, 1990). In oscillations, picture repetition has been found to result in a decrease in induced gamma band power (Gruber et al., 2004). While our repetition difference between conditions does not appear to be of concern for the theta band analysis, ERP signatures associated with repetition differences clearly overlap in time and are opposite in polarity to the N2 (and LPC) components that we expect as a result of our interference manipulation. Having two naming rounds should enable us to disentangle the two: repetition differences should disappear after the first round of naming, and should no longer affect the second round.

Participants were asked to name pictures in Italian to the best of their knowledge. The timings were identical to those in the picture naming tasks during the interference phase. The experimenter coded answers for correctness and in doing so initiated the next trial. There was no time limit, and next to EEG data and accuracy, (delayed) naming latencies were recorded, measuring the time from question mark presentation to speech onset. The order of presentation of the pictures was again semi-random. Each participant got one of six lists from the pool of lists described for the Italian post-tests at the end of each learning day. We made sure that the final test list was different from both of these post-test lists for each participant.

3.2.4 | Response Coding and Behavioral Analysis

3.2.4.1 | *Accuracy Coding*

Because the majority of errors were partial productions (a participant saying 'albera' rather than 'albero'; 78% of errors; 3% of all data), participants' Italian word productions during the final test on day 3 were coded on the phoneme level. For each production, we counted the number of correctly and incorrectly produced phonemes (see Chapter 2 and de Vos et al., 2018). Incorrect productions could be either insertions, deletions or substitutions (see Levenshtein, 1966). Table 3.3 exemplifies the scoring procedure for the 'albera' example.

Table 3.3

Scoring example, phonetically transcribed.

Target word	α	l	b	e	r	o
Participants production	α	l	b	e	r	a
Scoring	correct	correct	correct	correct	correct	incorrect (substitution)

'Albera' would be counted as having five correct phonemes and one incorrect phoneme. Together these two numbers (5,1) formed the basis for the dependent variable for statistical modelling. For data visualization and to provide descriptive statistics, we additionally calculated an error percentage based on these two numbers. This percentage corresponds to the number of incorrect phonemes out of the total number of phonemes (e.g., for 'albera': (1/(5+1))*100 = 16.67%).

3.2.4.2 | *Naming Latency Coding*

Naming latencies were measured manually from question mark presentation until speech onset using Praat (version 5.3.78, Boersma, 2001). Note that they reflect delayed naming latencies, rather than immediate naming latencies.

3.2.4.3 | *Modelling*

All behavioral data were analyzed in R (Version 3.5.1, R Core Team, 2018) using the lme4 package (version 1.1-21, Bates et al., 2015). As in Chapter 2, accuracy data were analyzed using a generalized linear mixed effects model of the binomial family, fitted by maximum likelihood estimation, using the logit link function and the optimizer 'bobyqa'. The dependent measure for this analysis was the odds of correctly producing a phoneme for a given target word. A two-column matrix with the number of correct and incorrect phonemes for each target word was passed to the model as dependent variable (this is one of multiple ways of specifying the response variable in binomial models, see also: https://www.rdocumentation.org/packages/stats/versions/3.2.1/topics/family). We tested for main effects of Interference (two levels: no interference, interference) and Round (two levels: first round, second round), as well as for their interaction to see whether the interference effect differed in magnitude across rounds. Both fixed effects variables were effects coded (-0.5, 0.5), meaning that a negative estimate for Interference reflects lower accuracy rates for interfered compared to not interfered items, a positive estimate for Round reflects higher accuracy in round 2 than round 1, and a negative estimate for the interaction of the two would reflect a smaller interference effect in round 2 than round 1. Random effects were initially fitted to the maximum structure justified by the experimental design (Barr et al., 2013), which included random intercepts for both Subject and Item, as well as random slopes by Subject and Item for Interference and Round and their interaction. Random slopes were removed when their inclusion resulted in non-convergence to fit the maximum model justified by the data, or when they correlated with each other or their respective intercept above 0.95 to avoid over-fitting. The final models included only random intercepts for Subjects and Items as well as a random slope by Subject for Interference. All p-values were calculated by model comparison, using chi-square tests, omitting one factor at a time (while keeping the random effects structure constant).

Naming latencies were analyzed using a linear mixed effects model, fitted by restricted maximum likelihood estimation (using Satterthwaite approximation to degrees of freedom). Because we were interested in naming speed differences after the artificially introduced delay, we subtracted the 2000 ms delay from each naming latency before analysis. We then log-transformed those corrected latencies and ran the linear model on those log-transformed latencies. Fixed effects were the same as for the accuracy model and the random effects structure was also determined based on the same principles. In this model, a positive estimate for Interference reflects higher RTs for interfered than not interfered items, a negative estimate for Round reflects overall faster RTs in round 2 than 1, and a negative interaction would reflect a smaller interference effect in round 2 than 1.

For the analysis of EEG signatures during picture naming in the interference phase, we additionally calculated median splits for each participant based on their naming latencies for the interfered items during the first round of the final test in Italian. We used the naming latencies of the first round because this round reflects the cleanest measure of interference strength. This choice was further reinforced by the fact that we observed a trend towards an attenuation of the interference effect in RTs from round 1 to round 2 (see section 3.3.2.2.2).

3.2.5 | EEG Recording and Analysis

3.2.5.1 | *EEG Recording*

Continuous EEG was recorded from 57 active Ag-AgCl electrodes embedded in an elastic cap, following the international 10-20 system (ActiCAP 64ch Standard-2, Brain Products), as well as from an electrode placed on the forehead (serving as ground). EEG signals were referenced on-line to the left mastoid and re-referenced off-line to the averaged activity over both mastoids. Eye movements were recorded from a bipolar montage consisting of electrodes placed above and below the right eye, as well as electrodes on the left and right temples. Mouth EMG was measured with two electrodes next to the upper and lower right lip to later on be able to tell when participants talked. All data were amplified with a BrainAmp amplifier, digitized with a 500 Hz sampling rate and filtered online with a high cutoff at 125 Hz and a low cutoff at 0.016 Hz. Impedances for EEG electrodes were kept below 15 kΩ.

3.2.5.2 | *EEG Preprocessing*

All off-line EEG processing was done using the Fieldtrip toolbox (Oostenveld et al., 2011) in Matlab (2018b, The Mathworks Inc.). The EEG signal was re-referenced to the average activity over both mastoids, low-pass filtered at 40 Hz, segmented into epochs from 500 ms before until 1500 ms after picture presentation, and detrended using the entire epoch. Trials containing artifacts, such as blinks or muscle activity, within the time window for analysis (-200 to 1000 ms after picture presentation) were removed. Eye blinks were identified using the EOG artifact detection function implemented in Fieldtrip. In addition, trials with amplitudes below -100 µV or above 100 µV, or peak-to-peak activity greater than 150 µV were discarded. These exclusions resulted in a total loss of 8% of the data.

3.2.5.2.1 | ERPs

For the analysis of event-related potentials, in line with previous research, the data were furthermore baseline-corrected based on the average EEG activity in the 200 ms interval before picture presentation. We subsequently averaged EEG activity for each participant across trials for each of the interference conditions.

3.2.5.2.2 | Oscillations

For the analysis of oscillatory power differences between conditions, we first computed time frequency representations (TFRs) of power for each of the conditions. TFRs were computed time-locked to picture presentation onset at frequencies ranging from 2 to 30 Hz, using a sliding window of three cycles, advanced in steps of 10 ms and 1 Hz. The data in each time window was multiplied with a Hanning taper, and subsequently Fourier-transformed. To test for an effect of interference condition, we subsequently calculated the difference between conditions per participant relative to the average activity in both conditions for that participant. The difference was calculated such that a positive difference reflects more power for interfered compared to not interfered words. This normalization of the condition differences made additional baseline correction unnecessary. Using cluster-based permutation tests, we compared this difference between conditions to zero (i.e., to the null hypothesis that there are no differences between conditions).

3.2.5.3 | *EEG Analysis*

EEG data were assessed inferentially using nonparametric cluster-based permutation tests (Maris & Oostenveld, 2007). This method allows for the statistical comparison of multi-dimensional (M)EEG data from two conditions while controlling for multiple comparisons, which arise when comparing multiple distinct data points (i.e., time-channel and channel-time-frequency data). The method first determines spatiotemporal or spatio-spectral-temporal clusters (that is clusters of adjacent time points and sensors, or adjacent time points, sensors and frequencies) that exhibit a similar difference across conditions. It does so by means of dependent-samples t-tests at each spatiotemporal or each spatio-spectral-temporal data point, thresholded at an alpha level of .05. Spatial adjacency was defined based on a neighbourhood structure in which channels had on average 6.5 neighbours. Each observed cluster's test statistic (the sum of all t-values contributing to it) was subsequently compared to a distribution of cluster statistics obtained through 2000 Monte-Carlo permutations based on random partitions of the data. P-values of the observed clusters were calculated as the proportion of these random partitions that resulted in a larger effect (i.e., a larger cluster statistic) than the observed effect. For tests with resulting *p*-values close to the critical alpha level of .05, we reran the analysis with 5000 permutations to obtain a more reliable Monte Carlo p-value estimate.

Using these cluster based permutation tests, we tested for differences between interfered and not interfered items at final test in Italian, both in ERPs and in oscillations. For both analyses, we first tested for an interaction of Interference (interfered vs. not interfered words) and Round (1st and 2nd round of final test). To do so, and following the procedure detailed in the Fieldtrip tutorial documentation, we first calculated the averaged difference between the two interference conditions (interference – no interference) for each person and for each of the two rounds. We then statistically compared the two resulting difference structures (one for each round) via a permutation test using a dependent samples t-test. A significant difference between condition differences for the two rounds reflects a significant interaction effect. Significant interactions were followed up with separate permutation tests for each of the two rounds of the final test in Italian, whereas non-significant interactions were followed up by an analysis of both rounds of the final test combined.

For the data from the first and last rounds of the picture naming task in the interference phase, we opted to analyze the two rounds separately without conducting an interaction analysis first. Our hypothesis applied most clearly to the first round of picture naming, as explained in section 3.3.3.2 in more detail, and the small sample size due to the median split approach was not suited for an interaction analysis.

3.2.5.3.1 | ERPs

We hypothesized to find differences between conditions in the amplitude of the N2 component, and hence ran targeted permutation tests in a restricted time window from 200 to 350 ms. In addition to that, we also ran exploratory permutation tests for a later time window (350-1000 ms), which encompasses the LPC.

3.2.5.3.2 | Oscillations

Based on previous research, we restricted the permutation tests for the time-frequency domain to the theta frequency band (4 -7 Hz). In line with the literature on theta effects described in the Introduction (section 3.1.2.2), we tested for theta differences in a window from 0 until 1000 ms after picture presentation and over the entire scalp.

3.3 | Results

3.3.1 | Exclusions

3.3.1.1 | *Exclusions from Accuracy Analysis*

For analysis of the behavioral accuracy data during the final test in Italian, we excluded words that a participant already knew in Italian before starting the experiment (as established in the familiarization task on day 1, 1% of data), words that he/she did not manage to learn in Italian (as established in the Italian post-test on day 2, 4% of data), as well as words they did not know in English (as established in the familiarization task during interference on day 3, 2% of data). In total these exclusions resulted in 6% of data loss (*M* = 6%, *SD* = 6%, range = 0-22%), hence leaving for analysis, on average, 32 out of 35 trials per round in the interference condition and 33 trials per round in the no interference condition (the maximum per round and condition being 35).

3.3.1.2 | *Exclusions from Naming Latency Analysis*

On top of the exclusions for the accuracy analysis, from the latency analysis we additionally excluded trials in which participants were unable to name a picture or named it incorrectly during the final test in Italian (i.e., errors, 4% of data). We furthermore excluded trials on which participants took multiple attempts to name a picture correctly, as well as trials on which they responded too early, that is during the two seconds delay window (10% of data). After all these exclusions, we were left with, on average, 29 trials per round in the interference condition and 31 trials per round in the no interference condition.

3.3.1.3 | *Exclusions from EEG Analysis*

For the EEG analysis, we excluded all trials that were also excluded from the accuracy analysis, as well as trials with EEG movement artifacts, as described in section 3.2.5.2. Artifact rejection resulted in the loss of 8% of data. After all exclusions, we had, on average, 30 and 29 trials in the interference condition in round 1 and 2 respectively (range = 24 – 35), and 31 and 30 trials in the no interference condition in rounds 1 and 2 (range = 23 – 35). Note that we did not discard trials based on their naming performance at final test in Italian: that is, unlike for the naming latency analysis, we included trials with errors in the EEG analyses, as well as trials in which participants took multiple attempts at naming or named a picture too early (as long as this was after the critical analysis window, i.e., after 1000 ms post picture presentation, and hence did not contaminate the EEG signal). We include those trials because the EEG analyses reflect the activity in response to stimuli and are not conditional on the final response.

The same exclusion criteria held for the analysis of the interference data. Here we were left with an average of 15 and 14 trials for the low and high RT groups in the first round of picture naming during interference, and an average of 15 and 14 trials for the same groups in the last round. Cell sizes for these comparisons are smaller than for the final test, because these comparisons rely on fewer total possible trials (i.e., max. 18 trials per median split group).

3.3.2 | Behavioral Results

3.3.2.1 | *Learning Performance in Italian*

After the first Italian learning session on day 1, participants had learned on average 46 out of the 70 words (*M* = 46.48, *SD* = 10.85, range = 26-69). Learning success on day 1 was comparable between the two interference conditions (Interference: *M* = 66%, *SD* = 17%, range = 26-97%; No interference: *M* = 67%, *SD* = 16%, range = 40-100%). After the second Italian learning session on day 2, participants had learned on average 67 out of 70 words (*M* = 67.37; *SD* = 3.56. range = 56-70). Learning success was again equal for both interference conditions (Interference condition: *M* = 96%, *SD* = 5%, range = 83-100%; No interference condition: *M*=96%, *SD* = 6%, range = 77-100%), and overall very high.

3.3.2.2 | *Retrieval Performance in Italian after Interference on Day 3*

3.3.2.2.1 | Naming Accuracy

Mean error rates for the interfered and the not interfered items during final test in Italian are shown in Figure 3.2 and the corresponding model output is reported in Table 3.4. We observed a main effect of Interference in the predicted direction. Participants made more phoneme production errors on interfered compared to not interfered words. In model terms, this main effect is reflected in a negative estimate, because we model accuracy rather than errors, and phoneme production accuracy for target words is lower in the interference condition than in the no interference condition. There was also a main effect of Round, such that participants improved and made less errors overall from round 1 to round 2. Round, however, did not modulate the main effect of interference. The interference effect in accuracy / error rates was thus stable across the two rounds of the final test.

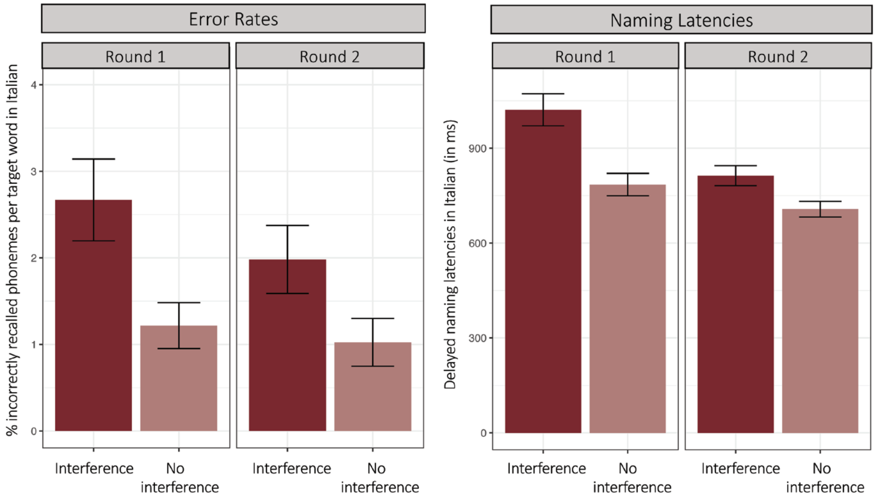

Figure 3.2
Error rates and delayed naming latencies during the final test in Italian on day 3. Error rates are expressed in the number of incorrectly produced phonemes per target word, and naming latencies reflect the time it took participants to name a picture after a 2 s delay period. Error bars reflect the standard error around the condition means.

3.3.2.2.2 | Naming Latencies

Mean naming latencies for interfered and not interfered items are shown in the right panel of Figure 3.2, and corresponding model outcomes in Table 3.5. We observed a main effect of Interference, such that interfered words took participants longer to recall than not interfered words. We also found a main effect of Round: participants were overall faster in round 2 compared to round 1 of the Italian final test. The interference effect was numerically bigger in the first round, the corresponding interaction term, however, did not reach statistical significance, indicating that the interference effect was present in both rounds and did not differ significantly in magnitude between rounds. Follow-up models for each round separately confirm this (round 1: $\beta = 0.15$, $t = 5.33$, $p(\chi 2) < .001$; round 2: $\beta = 0.08$, $t = 2.96$, $p(\chi 2) = .006$).

Table 3.4

Mixed effects model output for naming accuracy in the final Italian test on day 3.

Fixed effects	Estimate	*SE*	*z*	$p(\chi^2)$
Intercept	5.91	0.36	16.44	**<.001**
Interference	-0.77	0.31	-2.44	**.026**
Round	0.27	0.11	2.54	**.015**
Interference * Round	0.20	0.21	0.91	.389
Random effects	**Groups**	**Var**	***SD***	**Corr**
Item	Intercept	2.91	1.71	
Subject	Intercept	1.75	1.32	
	Interference	1.62	1.27	0.08

Note. Significant effects are marked in bold. *SE* = standard error; $p(\chi^2)$ = Chi-square p-value; Var = variance; *SD* = standard deviation; Corr = correlation.

Table 3.5

Mixed effects model output for log-transformed naming latencies in the final Italian test on day 3.

Fixed effects	Estimate	*SE*	*t*	$p(\chi^2)$
Intercept	6.47	0.05	125.51	**<.001**
Interference	0.12	0.02	4.94	**<.001**
Round	-0.08	0.02	-4.57	**<.001**
Interference * Round	-0.07	0.04	-1.92	.055
Random effects	**Groups**	**Var**	***SD***	**Corr**
Item	Intercept	0.03	0.18	
Subject	Intercept	0.06	0.24	
	Interference	0.01	0.08	0.77

Note. Significant effects are marked in bold. *SE* = standard error; $p(\chi^2)$ = Chi-square p-value; Var = variance; *SD* = standard deviation; Corr = correlation.

3.3.3 | EEG Results

3.3.3.1 | *EEG – Final Test in Italian*

Grand-averaged ERPs for the interfered and not interfered words during rounds 1 and 2 of the final test in Italian are shown in Figure 3.3. A time-frequency representation of the difference in induced activity between interfered and not interfered words is shown in Figure 3.4.

3.3.3.1.1 | ERPs – N2 Time Window (200-350 ms)

An initial permutation test revealed a significant interaction between Interference and Round (p = .002). This interaction was most prominent in an interval from 212 to 350 ms. Subsequent separate permutation tests for each of the two rounds of the final test revealed a large positivity for interfered compared to not interfered words in the first round (p = .001). This effect was most prominent between 204 and 350 ms and over centro-posterior electrodes. Visual inspection reveals that this positivity is best described as an attenuated negativity for interfered compared to not interfered items (see Figure 3.3). The direction of the effect and its scalp topography suggest that it reflects the beginning of an attenuated N400 for more recently seen pictures (i.e., the interfered items) compared to less recently seen pictures (i.e., the not interfered items). A follow-up analysis on a time window encompassing the classical N400 effect (200-500 ms) confirms this. The permutation test again revealed a significant positive shift (or in other words, a less negative shift) for interfered compared to not interfered items in this window (p = .002), which was most prominent over centro-posterior electrode sites and between 204-428 ms.

In the second round, we instead observed the expected N2 modulation. The permutation test revealed a larger negativity for interfered compared to not interfered items (p = .019). This difference between conditions was most pronounced in a time window from 218 to 316 ms and over frontal electrodes, which coincides well with the typical time course and topography of the N2. The ERP signatures in this early time window thus reverse from round 1 to round 2. The N2 effect in the second round confirms our hypothesis and the reversal of signatures suggests that recency differences between items were successfully eliminated after the first round of naming.

3.3.3.1.2 | ERPs – Later Time Window (350-1000 ms)

The interaction term between Interference and Round from 350 to 1000 ms post picture presentation did not reach statistical significance (p = .061). A follow-up permutation test over both rounds of the picture naming test together revealed a wide-spread negative cluster for interfered compared to non-interfered items (p = .007). Visual inspection of the grand average revealed that this cluster reflects a late positive component (LPC), that is attenuated for the interfered items as compared to the not interfered items, most pronounced from 428-636 ms. This LPC is present in both rounds (see Figure 3.3 for grand averages and cluster plots for each round separately).

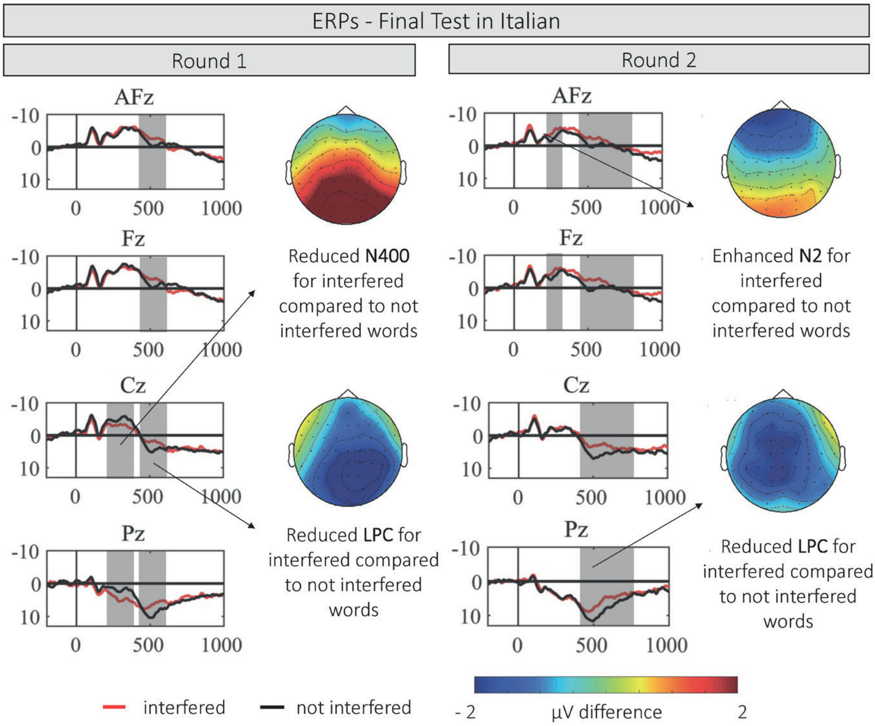

Figure 3.3
Grand-averaged ERP waveforms for interfered and not interfered items during rounds 1 and 2 of the final Italian picture naming test. Significant clusters revealed by the permutation tests are marked in grey. For each cluster a topographic plot is included. Colors indicate the amplitude difference (in µV) between interfered and not interfered items, such that shades of red reflect more positive going ERPs for the interfered compared to the not interfered items, and shades of blue reflect more negative going ERPs for interfered items.

3.3.3.1.3 | Oscillations – Theta Band (4-7 Hz)

In the time-frequency domain, there was no significant interaction between Round and Interference ($p = 1$). A follow-up permutation test of the data collapsed over both rounds of the final naming test revealed a large cluster in the theta frequency band ($p = .004$). Retrieval of interfered items thus resulted in more induced theta activity than retrieval of not interfered items, which we observed most prominently in a time interval of 510-1000 ms and over the entire scalp.

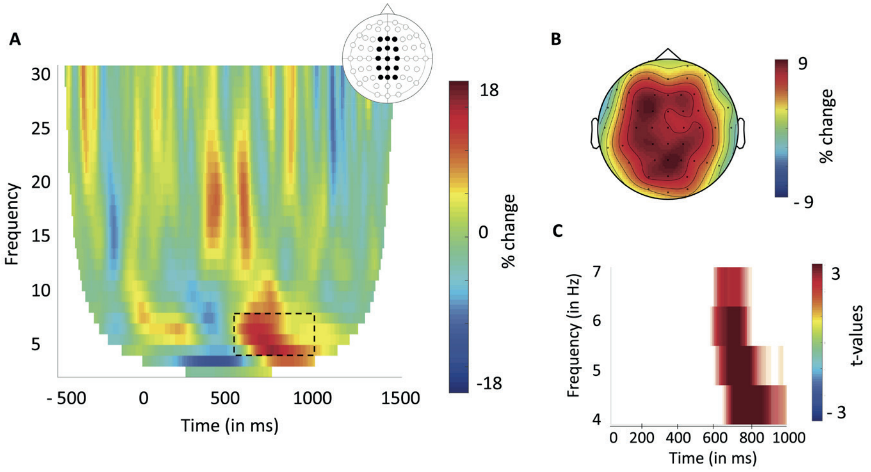

Figure 3.4
A. Time-frequency representation of power differences between interfered and not interfered items, averaged over a representative sample of channels involved in the cluster revealed by the permutation test (see black dots in the topoplot in the right upper corner: Fz, F1, F2, Cz, C1, C2, FCz, FC1, FC2, CPz, Cp1, Cp2, Pz, P1, P2). Shades of red reflect more theta for interfered compared to not interfered items. Power differences were calculated relative to the average activity in both conditions, and thus reflect a percent power change. Dashed lines reflect the significant cluster. **B**. Scalp distribution of power changes for the interference condition minus the no-interference condition (relative to the average activity in both condition), averaged from 510-1000 ms and for frequencies between 4-7 Hz. **C**. Statistical map for the theta effect in time and frequency, averaged over the same channels as in A. Colors reflect t-values.

3.3.3.2 | *EEG – Interference Phase in English – Interfered Items Only*

To test whether activity during the interference phase was directly related to retrieval performance at final test, we analyzed the interference phase data conditional on participants' naming latencies at final test in Italian. Based on a median split, we divided each participant's interfered items into those that took participants long to recall at final test in Italian and those that took them relatively less long to recall. The inhibitory control account of forgetting would attribute such retrieval difficulty discrepancies to differences in inhibition during the interference phase: Italian labels that are more difficult to recall at final test must have been inhibited more during retrieval of their English translation equivalents in the interference phase. If competition and inhibition during the interference phase are indeed responsible for the retrieval difficulty differences between items at final test, we should thus observe a higher amplitude N2 and more theta power for items that are slow to retrieve at final test (high interference group) compared to items that are fast to retrieve at final test (low interference group). This hypothesis concerns most directly the first round of picture naming in the interference phase. We speculate though that by the last round, differences between the two conditions disappear. To that end, we analyzed grand averages for both the first and fourth (i.e., last) round of picture naming during interference.[7] Grand averages and topoplots contrasting the two median split interference groups are shown in Figure 3.5.

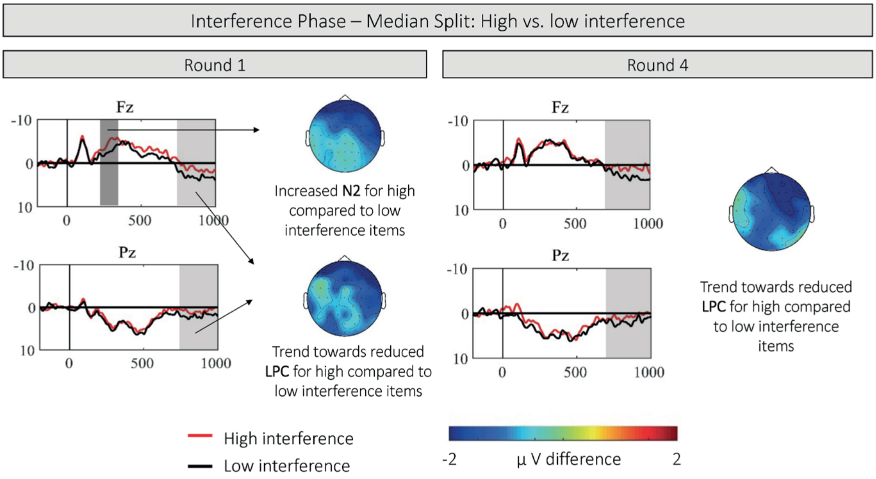

Figure 3.5
Grand-averaged ERPs for items from the high and low interference groups (as determined through a median split of naming latencies from the final Italian test) from the English interference naming task, rounds 1 and 4. Topographies of significant effects and trends in the data are displayed to the right of their respective grand averages. Colors indicate the amplitude difference (in μV) between conditions, such that shades of blue reflect more negative going waveforms for highly interfered items compared to less interfered items.

3.3.3.2.1 | ERPs – N2 Time Window (200-350 ms)

In the first round of picture naming during interference, we indeed observed a larger N2 for highly interfered items (i.e., items that later took relative long to produce in the final Italian test) compared to less interfered items (p = .049). The difference between conditions was most pronounced over frontal electrodes and between 218

[7] Note that because the second block of picture naming was preceded by two rounds of phoneme monitoring in English, round 4 of picture naming corresponds to round 6 of the interference phase overall. Rounds 1 and 4 are thus separated by 5 intermittent retrievals rather than just 3 (as their names might suggest).

and 350 ms post picture presentation. In the last round of picture naming, this N2 was no longer present (i.e., no significant clusters, *ps* = 1).

3.3.3.2.2 | ERPs – Later Time Window (350-1000 ms)

In the later time window, visual inspection suggests that there is a small late positive shift for high compared to low interference items both in the first and the fourth round of picture naming during interference. These differences, however, were not statistically robust (1st round: p = .066, differences most pronounced between 730-1000 ms; last round: p = .051, differences most pronounced between 728-1000 ms). These positive components differ from the LPC reported in the final test both in their temporal as well as their spatial distribution.

3.3.3.2.3 | Oscillations – Theta Band (4-7 Hz)

There were no significant differences between high and low interference items in the time-frequency representations of either of the two rounds of picture naming in the interference phase.

3.4 | Discussion

The present study aimed at unravelling the neural correlates of foreign language attrition. In Chapter 2, we had postulated that foreign language forgetting is the consequence of competition and inhibition between translation equivalents. Here, we asked whether we can track those processes on the neural level. To that end, participants first learned a set of new L3 Italian words over two consecutive days. On a third day, we interfered with their knowledge of these recently learned words by having them repeatedly retrieve half of the words in L2 English. Finally, we assessed the effect of this interference phase in a final recall test on all originally learned Italian words, also on day 3. Next to asking whether we can see neural evidence for competition and inhibition at final test, we also asked whether behavioral performance at final test can be related to the degree to which these processes are recruited during interference.

Behaviorally, we replicated Chapter 2. Participants were slower and less accurate in recalling Italian words that had been interfered with (i.e., named in English) than words that had not. In the EEG, these interference effects were accompanied by more theta power, an enhanced N2 and a reduced LPC for interfered compared to not interfered items. Moreover, differences in performance at final test went along with amplitude differences in the N2 component during the interference phase. We report an enhanced N2 for items that took participants long to recall at final test compared to items that were easier to retrieve and hence interfered with less successfully. Together, these findings establish the N2, the LPC and oscillatory power in the theta band as neural correlates of foreign language attrition.

The replication of the behavioral interference effects reported in previous lab-based language attrition studies (e.g., Bailey & Newman, 2018; Chapter 2; Levy et al., 2007) with a new language combination and a larger set of to be learned words confirms the robustness of these effects. They occur despite the fact that the learning phase in our experiment was spread over two days and that the reaction times at final test were measured after a delay rather than immediately after picture presentation.[8] The neural signatures that accompany these behavioral effects resemble those reported in various other strands of literature, including research on bilingual language production and forgetting more generally. Departing from how they are typically interpreted in these other areas, our EEG results provide converging evidence for the assumption that (foreign) language attrition can be the consequence of competition and interference from the more recent use of other languages.

3.4.1 | The N2 as a Marker for Interference-Induced Foreign Language Attrition

The frontal N2 component that we report for interfered compared to not interfered items in the second round of the final test resembles the N2 that is often found in language switching studies. In those studies, participants typically alternate between naming pictures in L1 and L2, and the N2 is found to be strongest on switch compared to repeat trials (particularly when a switch is made from L1 to L2; Jackson et al., 2001; Zheng et al., 2020). In line with reports of the N2 as a marker of response conflict and inhibition in non-linguistic tasks (e.g., Folstein & van Petten, 2008), language switching studies typically interpret their results as evidence for interference from and inhibition of a non-target language (e.g., the L1 when switching to naming pictures in L2; see Kroll et al., 2008, for a review). Observing a comparable N2 for interfered items at final test is thus compatible with the idea that between-language interference is (at least partially) responsible for the behavioral forgetting effects measured at final test. Specifically, it is in line with the proposal put forth in Chapter 2 that retrieval of interfered L3 words is hindered by competition from the recently practiced L2 words and that this interference is not (or much less) present for L3

[8] Note that in Chapter 2, raw naming latencies at final test were fairly long even without a production delay (roughly 2200 and 1700 ms for interfered and not interfered words respectively, see Appendix A.6). The delay of two seconds we introduced here for safety reasons (to avoid movement artifacts in the EEG) was thus not much of a delay for the average participant and hence unlikely to wash out retrieval speed differences between conditions.

words whose L2 translations were not recently retrieved. Whether the N2 reflects only the presence of this response conflict (i.e., interference between English and Italian labels), or in fact the active inhibition of the English competitors to allow for successful retrieval of the Italian words, is unclear. Either way though, the N2 provides corroborating evidence in favor of the idea that language forgetting can be caused through interference from recently retrieved translation equivalents.

Our N2 is comparable to the switching N2 both in terms of latency (200-350 ms post stimulus presentation) and scalp topography (fronto-central). This is interesting and, in fact, not trivial, because our study differs from mixed language switching studies in a number of ways. As explained in the Introduction (section 3.1.2.1), these differences include the timing of the switch (immediate vs. delayed), the level at which interference / inhibition is thought to act (language global vs. local, item-specific), and the languages involved (L1/L2 vs. L2/L3). We are only aware of three EEG studies that have addressed long-term switch effects and that tested item-specific switching on top of global switch effects (Branzi et al., 2014; Misra et al., 2012; Wodniecka et al., 2020). Using a blocked language switching paradigm, these studies ask what an entire block of naming in one language does to naming in another language in a subsequent block. Crucially though, these studies have no initial learning or familiarization phase, and hence compare naming in L2 after L1 with naming in L2 after *no* prior naming. Not surprisingly so, they report facilitation for naming a picture in L2 if the same picture had previously been named in L1 compared to when it had not been named before at all. Their behavioral effects, and hence their EEG effects, are thus not compatible with those reported here or in Chapter 2 or by Bailey and Newman (2018). The current study thus differs from both mixed and blocked language switching studies in important ways. That we nevertheless report a comparable N2 effect is in line with the idea that similar inhibition and interference mechanisms are at work in language switching and L2-induced L3 attrition. Just as global switching from naming pictures in L1 to naming pictures in L2 invokes an N2, so does the retrieval of words in Italian after a remote block of naming the same items in English.

3.4.2 | Oscillatory Theta Power as an Index of Between-Language Competition

In the frequency domain, we report more theta power for interfered compared to not interfered words at final test in Italian. Though different in terms of scalp topography, our theta effect fits with reports of interference-induced theta activity in other domains, such as, for instance, the non-linguistic cognitive control literature. In a go/no-go task, for example, mid-frontal theta power is typically higher on no-go trials, where the tendency to press a button needs to be suppressed, compared to go trials (e.g., Nigbur et al., 2011). Very similar to the N2, theta is hence understood to index the presence of a response conflict and possibly the recruitment of cognitive control processes to overcome this conflict. Next to the cognitive control literature, memory research on so-called retrieval-induced forgetting (RIF) effects has also consistently reported modulations in the theta band. These studies reported higher mid-frontal and left parietal theta power in competitive compared to uncompetitive retrieval situations, suggesting that theta indexes the amount of competition and thus interference that is encountered during item recall (e.g., Staudigl et al., 2010; Hanslmayr et al., 2010). Our theta effect is not restricted to mid-frontal or left-parietal electrode sites, and is instead more wide-spread. This topography difference is most likely attributable to differences in stimuli and task design between our experiment and the theta studies in other domains. Competition from translation equivalents and the suppression of a non-target language word likely requires a different kind of control than the suppression of a 'Go' response in a no-go trial or the suppression of semantic competitors in RIF paradigms. Remember also that some of the RIF studies compare two different tasks (e.g., active retrieval vs. passive restudy in Staudigl et al., 2010) and that the scalp topography of theta activity reported in these studies might thus also partially reflect differences in task design between the two conditions rather than interference alone, making it difficult to compare to our theta effect.

Regardless of the topography differences, we think that it is justified to conclude that the theta effect in our study reflects interference of a non-target language (i.e., English) during productive recall of words in a target language (i.e., Italian). Just as the N2 discussed above, the theta effect at final test thus corroborates the idea that between-language competition is at least part of the reason for why interfered Italian words at final test are less well recalled. To our knowledge, we are the first to provide evidence for increased theta power as a marker of between-language interference.

3.4.3 | The Consequences of Language Interference – the LPC

In the final test, next to the N2 and theta effects, we additionally observed a late positive component (LPC), reduced in magnitude for interfered compared to not interfered items in both rounds at final test. Both in terms of its central scalp distribution and latency (roughly 400-600 ms post stimulus onset), this signature is reminiscent of a similar late positive component in the memory literature. The 'memory' LPC is most typically found in studies on recognition memory, where it is stronger at retrieval for previously studied ('old') compared to previously unstudied ('new') items, and especially for items for which participants additionally make correct as compared to incorrect source judgments (i.e., recalling details of the original learning context; Rugg et al., 1995; Wilding, 2000). Its amplitude has furthermore been found to vary with decision certainty, such that it appears to be larger for items that people report to confidently remember as compared to items for which people only report a vague sense of familiarity (Smith, 1993). Given the conditions that elicit this component, the LPC is generally understood as a marker of conscious recollection success, and possibly an index of the quality of the information that is retrieved from episodic memory.

Though not specifically predicted, our finding of an enhanced LPC for not interfered compared to interfered items fits well with this recollection success interpretation. Memory representations of Italian labels in the no interference condition have not been interfered with and so retrieval for those items is easier, faster and ultimately more successful (as also seen in reaction times and error rates) than for interfered items. It thus seems plausible that the LPC in our study indexes retrieval success in Italian. Note that one could have also predicted the opposite pattern: a larger LPC for the interfered items because their corresponding pictures have been repeated more recently (Bentin & McCarthy, 1994). That this was not the case reinforces the interpretation that the LPC in our study indexes recollection processes specific to the Italian words, and not their associated concepts.

In the language domain, LPC effects have been found to index lexicality and conscious semantic access. Bakker et al. (2015b), for example, reported a reduced LPC for newly learned words (in L1) compared to existing words and partial evidence for an increase in the magnitude of the LPC with consolidation of these novel words. Their LPC effect, however, had a fronto-central scalp distribution and was furthermore elicited under very different task demands (semantic relatedness judgments between the words and unrelated primes), and is hence difficult to compare directly to our findings. Even though the comparison is not straightforward, if our LPC were to index degree of lexicality, this would mean that words in the interference condition, despite having been learned to the same criterion as not interfered words, lack behind in lexicalization, or that their lexical representations have undergone erosion due to interference. In a follow-up experiment, it would be interesting to establish degree of lexicality (i.e., LPC amplitude) prior to interference, to see exactly what changes interference brings about. Do interfered items decrease in lexicality (i.e., decrease in LPC magnitude) due to interference or do they simply stagnate, compared to not interfered items (i.e., LPC amplitude increases for not interfered items and remains the same for interfered items)?

Curiously, some of the mixed language-switching studies described earlier tend to report an LPC opposite to that in our study (i.e., larger for switch compared to repeat trials, e.g., Jackson et al., 2001). Not all language switching studies report an LPC though, making it unclear what the precise conditions for its emergence are. Most likely, the switching LPC reflects different processes than the LPC we report here and future research will be necessary to fully understand its functional significance in multilingual language production. Based on the present results, and the available evidence from other strands of research, we conclude that the LPC is a marker for retrieval success and as such reflects the consequence of between-language interference, namely reduced accessibility to interfered compared to not interfered Italian labels.

3.4.4 | Disentangling Recency from Interference

One aspect of the final test that warrants discussion is the fact that we observed the predicted N2 modulation only in the second round of the final test, whereas we did find effects in theta power and the LPC in both rounds. In place of the N2, we observed a reduced (rather than enhanced) negativity for interfered compared to not interfered items in the first round of the final test, which we interpreted as an attenuated N400 based on its latency and topography. This N400 most likely reflects recency differences between items in the two conditions. Though equally familiar initially, the pictures corresponding to the interfered items were seen more recently than those of the not interfered items, and hence were less surprising and easier to process, resulting in an attenuated N400 (Bentin & McCarthy, 1994). Differences between conditions caused by recency appear to be much stronger than differences due to interference and so the N400 (larger in amplitude for *not* interfered items) overwrote the N2 (larger in amplitude for interfered items) in the first round. By round two, recency differences between items had disappeared, enabling us to observe the predicted interference-related N2. In contrast, neither the LPC nor theta power appear to be influenced by such recency differences. In the frequency domain, previous literature only implicated the gamma frequency range in picture repetition (Gruber et al., 2004). The LPC, in turn, has been found to be sensitive to picture repetition, yet in the opposite way, being larger for repeated (i.e., interfered items in our study) compared to not repeated items (e.g., Bentin & McCarthy, 1994). The processes that our LPC effect reflects (i.e., recollection success for Italian labels) appear to have been stronger than item differences due to picture repetition.

While this confound is unfortunate, we would like to stress that recency differences are inherent to the design of our study. Eliminating them would require inclusion of the no interference items in the interference phase, in a task that does not require competitive retrieval of these words, but nevertheless exposes participants to their images. One could argue that we could have used a simple passive exposure task, akin to the EEG RIF studies mentioned earlier. However, given that our stimuli are meaningful words, relevant not only within the context of the experiment itself, it is very possible that even in a passive exposure condition (or in fact in any task), participants would covertly retrieve the words (in whatever language). Such word retrieval would have interfered with our experimental manipulation in that the words from the no interference condition would then also have received interference. To weaken recency differences, future studies might want to consider using different pictures in all experimental phases. All pictures would then be equally new at final test and differences in ease of visual recognition would no longer be contaminating the signal. Note though that items in the interference phase would still be *conceptually* more recent and might thus still be easier to access even with a different set of pictures (see also Chapter 6 for a discussion of this aspect). The latter risks and considerations are why we instead stayed with the piloted and established paradigm used in Chapter 2.

3.4.5 | Linking Activity During Interference to Later Forgetting

So far, we have looked at EEG activity during final recall of Italian items and found evidence for competition and interference at that moment (theta and N2) as well as the immediate consequences of this interference for recall success (LPC). While competition and interference at final test suffice to explain the observed behavioral forgetting effects, interference-driven (language) forgetting is typically assumed to already be induced during the preceding interference phase (Anderson, 2003; Chapter 2). Studies on the neural correlates of retrieval-induced forgetting support this claim (e.g., Johansson et al., 2007; Hanslmayr et al., 2010). Staudigl et al. (2010), for example, found that participants who showed the greatest decrease in theta activity over multiple rounds of competitive retrieval (in the interference phase) also forgot more of the very competitors that caused the competition during retrieval. Staudigl and colleagues interpret the competition reduction that takes place across subsequent rounds of retrieval to reflect the amount of inhibition that is applied to competitors. The more inhibition is applied, the more troublesome retrieval is for those competitors at subsequent final test, and hence the larger the forgetting effect.

Here, we asked whether a similar relationship between activity during the interference phase and final test also holds for the language case. Our median split analysis of the interference phase data reflects a first step towards understanding the temporal dynamics of interference-induced foreign language attrition. We split each participant's items into high and low interference items depending on how fast they were recalled at final test. Items that took a participant relatively long to recall at final test must have been interfered with more than items that were faster to recall at final test. The former should hence show more evidence for interference (and possibly inhibition) during the interference phase than the latter, if there is a direct relationship between the two experimental phases. While we did not observe a modulation of theta power during the interference phase, we did find differences between the two types of items in the amplitude of the N2 component. In the first round of picture naming during the interference phase, we observed a higher N2 amplitude during English retrieval of items that were subsequently more difficult to retrieve in Italian than items that were relatively easy to retrieve at final test. There is thus indeed a quantifiable relationship between activity during the interference phase and later retrieval ease. Assuming that the N2 reflects the presence of interference from response alternatives (i.e., Italian labels during English picture naming) and possibly the need for inhibition of those competing responses for successful retrieval of the target response (i.e., the English label), the current pattern of results suggests that the extent to which Italian labels interfered and/or were inhibited is directly related to how well they were recalled at final test. The behavioral interference effects are thus not only the result of competition at final test, but are already set in motion during the preceding interference phase.

Interestingly, in the last round of picture naming during the interference phase, the N2 was no longer enhanced for highly interfered as compared to less interfered items, suggesting that retrieval differences at final test were induced at the beginning of the interference phase rather than later on. After multiple rounds of retrieval in English, the Italian translations in the high interference group no longer interfered more and no longer needed extra inhibition than items in the low interference condition. It should be noted though that this decrease was only descriptively observed in the current study. The small sample size did not allow for a statistical comparison of the two rounds of picture naming in the interference phase (i.e., no interaction analysis with round was possible).

We encourage future research to follow up on our interference phase analysis, not only to replicate the N2 findings, but also to better understand why neither theta power nor the LPC amplitude reliably distinguished later well and less well recalled items. As already noted, the interference phase analysis is based on a relatively small number of trials per condition (15 trials on average) and so it is possible that we simply did not have enough power to reliably detect theta power and LPC amplitude differences. A follow-up study with more items, and possibly without a no interference condition (allowing for all 70 learned Italian items to be part of the interference phase) would help disentangle the current pattern of results.

3.4.6 | Conclusion

The current study established the N2, the LPC and oscillatory power in the theta band as neural markers of foreign language attrition. Their presence at final test and (at least partially) during the interference phase supports the idea that foreign language forgetting is the result of competition dynamics between translation equivalents in multiple languages. At final test in Italian, oscillatory power in the theta band and the N2 component of the event-related potential reflected interference from (and possibly inhibition of) the recently practiced English translation equivalents. The LPC, in turn, based on its occurrence in the memory literature, most likely reflected the consequences of this competition between English and Italian labels and indexed the reduced accessibility to interfered compared to not interfered Italian labels. Finally, we were able to link activity during the preceding English interference phase to later retrieval speed in Italian. An enhanced N2 for items that were later most difficult to retrieve is in line with the idea that competition and inhibition during the interference phase are causally related to later retrieval ability at final test. Taken together, our results provide the first converging neural evidence for the idea that foreign language attrition can be caused by the more recent practice of words in another foreign language.

CHAPTER 4

New in, Old out: Does Learning a New Language Make You Forget Previously Learned Foreign Languages?

This chapter is based on:

Mickan, A., Slesareva, K., McQueen, J. M., & Lemhöfer, K. (2020). *New in, old out: Does learning a new language make you forget previously learned foreign languages?* Revised version of manuscript submitted for publication.

I would like to thank Panthea Nemat and Camí Geenen for help with data acquisition and coding for Experiment 1 and Katya Slesareva for help with data acquisition and coding for Experiment 2.

Abstract

Anecdotal evidence suggests that learning a new foreign language (FL) makes you forget previously learned FLs. To seek empirical evidence for this claim, we tested whether learning words in a previously unknown L3 hampers subsequent retrieval of their L2 translation equivalents. In two experiments, Dutch native speakers with knowledge of English (L2), but not Spanish (L3), first completed an English vocabulary test, based on which 46 participant-specific, known English words were chosen. Half of those were then learned in Spanish. Finally, participants' memory for all 46 English words was probed again in a picture naming task. In Experiment 1, all tests took place within one session. In Experiment 2, we separated the English pre-test from Spanish learning by a day and manipulated the timing of the English post-test (immediately after learning vs. one day later). By separating the post-test from Spanish learning, we asked whether consolidation of the new Spanish words would increase their interference strength. We found significant main effects of interference in naming latencies and accuracy. Participants sped-up less and were less accurate to recall words in English for which they had learned Spanish translations compared to words for which they had not. Consolidation did not significantly affect these interference effects. Learning a new language thus indeed comes at the cost of subsequent retrieval ability in other FLs. Such interference effects set in immediately after learning and do not need time to emerge, even when the other FL has been known for a long time.

4.1 | Introduction

Most multilinguals share the intuition that learning (words in) a new foreign language (L3) comes at the cost of retrievability of (words in) previously learned foreign languages (L2). Studies on third language acquisition have paid surprisingly little attention to these frustrating side effects of learning a new language.[9] In fact, to date, there is little experimental evidence documenting them and there have been few attempts to provide an explanation for why learning a new language might negatively affect previously learned ones (though see research on the bilingual disadvantage, e.g., Gollan & Silverberg, 2001; Gollan et al., 2005). The present study seeks to fill this research gap and asks whether learning a new foreign language indeed hampers accessibility to other foreign languages in the very early stages of vocabulary learning. It also aims at providing insights into when, and thus why, such effects emerge.

A possible explanation for why a new foreign language might interfere with older ones is language competition. The languages of a multilingual are thought to interact and compete with one another (Kroll et al., 2008). When a Spanish-English bilingual wants to refer to a ‹table› in English, the Spanish word ‘mesa’ will be activated along with the target English word. This co-activation can result in between-language lexical competition, which in turn can delay selection and hence production of the target language word, and in extreme cases even lead to complete retrieval failure (e.g., Colomé, 2001; Hermans et al., 1998; Kleinman & Gollan, 2018). Based on these documented online language competition effects, it has been proposed that between-language competition can, in the long run, lead to language forgetting (i.e., attrition). In Chapter 2, for example, participants first learned a set of new L3 Spanish words. A day later, participants repeatedly retrieved half of these words in either L1 Dutch or L2 English. In a subsequent Spanish recall test on all originally learned words, recall proved less accurate and slower for Spanish words that had been interfered with (i.e., retrieved in L1/L2) compared to words that had not been interfered with. These interference effects persisted for an entire week, at least in reaction times, thus linking language competition to long-lasting changes in retrieval ease. Our findings from Chapter 2 are supported by a recent study from Bailey and Newman (2018), who showed that newly learned L2 Welsh words also take longer to retrieve after retrieval practice of their translation equivalents in L1 English. Together, these studies (and others, e.g., Isurin & McDonald, 2001) clearly point towards a role for between-language competition in language forgetting.

[9] Previous research has instead focussed on the opposite, namely on how already known languages (L1 or L2) affect the acquisition of a new L3 (e.g., Bardel & Falk, 2007).

Interestingly though, all of the above studies test for competition effects on relatively new, recently acquired foreign language material. It remains unclear whether 'old' memories, learned long before, are also affected by such competition dynamics. To understand the difference between just recently acquired words and long known words, it is important to make a distinction between episodic memory on the one hand, and semantic memory on the other (Baddeley, 2015). Episodic memory refers to contextual knowledge, such as memory for specific events, whereas semantic memory refers to abstract knowledge of meanings and facts, independent of the episodic context they were acquired in. Most traditional memory studies on forgetting make use of episodic memory tasks to test their theories (e.g., list learning that is only meaningful within the specific context of the experiment; Anderson et al., 1994; Müller & Pilzecker, 1900). The established mechanisms of forgetting, including forgetting by competition, have thus mostly been tested on episodic memories, leaving it unclear whether they equally apply to abstract semantic memory content. Lexical knowledge (i.e., a language's vocabulary) is an example of semantic memory. However, newly-learned words actually start out as episodic memory traces too and only get transferred to semantic memory through offline consolidation, a process that is assumed to start immediately after learning, but that can take up to multiple days to complete (Davis & Gaskell, 2009; Gaskell & Dumay, 2003). Given that the above studies on language attrition allow for very little, if any, consolidation of the new words (Bailey & Newman, 2018; Chapter 2; Isurin & McDonald, 2001), these studies again only demonstrate the impact of language competition for relatively fresh, and most likely still episodic memories, rather than established semantic knowledge, such as the words of an already known L2. It is conceivable that well-established old knowledge is less vulnerable to interference than the fresh L2 knowledge tested in the studies above. For this reason, testing for interference on words consolidated long ago, as will be done in the present study, is a much stricter test of the role of between-language competition during language attrition.

What is more, in previous studies on attrition, interference usually came from the repeated retrieval of known words in other languages rather than from the new learning of such words. Effects of new learning on previously learned material are commonly known as retroactive interference (RI) effects in the memory literature (Müller & Pilzecker, 1900). Usually though, studies investigating RI involve the learning of two (new) lists of items in *immediate* succession, thus not testing for the effect of new learning on 'old' semantic knowledge, as we propose to do in the current study. In typical RI experiments, the reason for impaired recall of items on the first learned list can be attributed to the fact that the second learned list interferes with the consolidation of the former (see Müller & Pilzecker, 1900, for the original formulation of this hypothesis). This type of interference is less applicable to foreign language forgetting, because it is hardly ever the case that two new languages are learned at the same time. In contrast, some RI studies that have allowed for (at least partial) consolidation of items in the first list before the learning of the second list suggest that retroactive interference has little impact on consolidated episodic knowledge (e.g., Ellenbogen et al., 2006; Landauer, 1974; Skaggs, 1925). Applying these insights to the language learning situation under investigation in the current study, one would thus expect little interfering effects of learning a new language on an already known foreign language, given that the latter has been consolidated and hence supposedly been rendered 'resistant' to interference.

The empirical evidence thus seems to speak against the anecdotal reports that learning a new foreign language interferes with existing memory for a previously acquired foreign language. In this chapter, we report on two experiments that test whether this conclusion is justified. More specifically, we ask if, and under which circumstances, new language learning hampers access to previously learned and well-consolidated foreign language words. In the first experiment, we tested a group of Dutch native speakers with good command of L2 English. Participants first completed a picture-based English vocabulary test, on the basis of which a set of participant-specific, *known* English words was chosen. For half of those words, their L3 Spanish translations were subsequently learned, and hence the English words were supposedly interfered with, while this was not the case for the other half. Finally, participants' picture naming accuracy and speed for all L2 English words was measured. If the learning of a new language comes at the cost of remembering already known languages, we should see longer naming latencies and – if this cost is as severe as has been shown for recently acquired L2 knowledge (Chapter 2) – more errors in English word productions for which Spanish translations were learned compared to words for which no translation equivalents were learned.

4.2 | Experiment 1

4.2.1 | Method

4.2.1.1 | *Participants*

Thirty-one Dutch native speakers with normal or corrected-to-normal vision and without a history of neurological or language-related impairments were recruited from the Radboud University participant pool. Four of them had to be excluded because they did not know enough English words in the pre-test to construct a matched item list (see section 4.2.1.2.1). One additional person had to be excluded due to a technical failure. The remaining 26 participants (16 female) were aged between 18 and 27 ($M = 21.77$) and had Dutch as their only mother tongue.

As determined via an online language background questionnaire completed before participants came to the laboratory, none of the participants, with the exception of one, had any knowledge of Spanish prior to the experiment. The one participant who did report having learned Spanish had just started doing so using a language learning app (Duolingo) three weeks prior to participating in the experiment, and judged their Spanish as very poor (1 out of 7 in all domains, i.e., reading, writing, listening, speaking). All participants reported English as their first and most frequently spoken foreign language. We asked for proficiency self-ratings on a Likert scale from 1 to 7 and their frequency of use of English per day in minutes. Both self-rated proficiency and frequency of use were assessed separately for the four language domains (speaking, listening, reading and writing). After the main experiment, we also measured participants' English vocabulary size using LexTALE (Lemhöfer & Broersma, 2012). The results of these measures are summarized in Table 4.1. Other foreign languages participants had learned included French, German and Latin. As in all previous chapters, we refer to Spanish as an L3 regardless of how many other foreign languages a participant had learned before the study. Participants gave informed consent and received either course credit or vouchers for participation (10€/h). The study was approved by the Ethics Committee of the Faculty of Social Sciences, Radboud University.

4.2.1.2 | *Materials*

The complete item set consisted of 103 nouns referring to concrete objects and animals (see Appendix C.1 for a full list). They were non-cognates between Dutch, English and Spanish and were one or two syllables long in English ($M = 1.33$, $SD = 0.47$) and between two and four syllables long in Spanish ($M = 2.59$, $SD = 0.66$). Lemma frequencies of the corresponding Dutch words ranged from 0 to 200 per million ($M = 35.90$, $SD = 49.48$) (Baayen et al., 1995). To match items in terms of frequency between the interference and no interference sets for each participant, we used the corresponding log frequencies, which ranged from 0 to 2.32 ($M = 1.08$, $SD = 0.63$). Pictures representing the nouns were photographs taken from Google images or the BOSS database (Brodeur et al., 2010) and were presented on a white background (occupying a maximum of 400 px in either width or length). Each Spanish noun was recorded spoken by a female Spanish native speaker from Andalucía (Spain).

Table 4.1
Participant characteristics.

	M	*SD*	*range*
English AoA	10.85	1.52	8-14
English LoE (years)	10.69	3.39	6-18
English Frequency of Use (min/day)			
Speaking	16	27	0-120
Listening	153	120	0-480
Reading	49	38	0-120
Writing	17	22	0-60
English Proficiency[a]			
Speaking	5.04	0.87	3-7
Listening	6.04	0.72	5-7
Reading	5.73	0.96	4-7
Writing	5.12	1.07	2-7
English LexTALE	74.89	12.04	51-95

Note. *M* = mean; *SD* = standard deviation; AoA = Age of acquisition; LoE = length of exposure (i.e., amount of years participants had been learning English). [a]Proficiency self-ratings were given on a scale from 1 (very poor) to 7 (like a native speaker).

4.2.1.2.1 | Item Selection

Per participant, 46 nouns were selected on the basis of the participant's pre-test results. Nouns had to be known in English, that is correctly named in the pre-test at first attempt. The first 46 words from the pre-test were considered the 'ideal' item set. If a participant did not know one or more words from this set, these were subsequently replaced with known words from the remaining pre-test items. A Matlab script (v.8.6, R2018b, The Math Works, Inc.) took care of the replacement and made sure that replacements were as similar as possible in terms of the item matching criteria (see below) to the original item from the base set (mean words replaced = 13.03; range = 1-30, see Appendix C.2 for details).

Importantly, each participant's final set of 46 items consisted of two subsets: 23 words that would be learned in Spanish (interference set) and 23 words that would not (no interference set). Words in the two subsets were matched on Spanish and English word length (measured in syllables), within- and across-set semantic similarity (via distance values derived from semantic vectors, small values indicate high similarity, zero being the value for two identical items, Mandera et al., 2017) as well as on phonological similarity with all other items in a participant's entire set, both in English and Spanish (Levenshtein distances based on letters, a small value indicates high similarity between two items; Levenshtein, 1966) and Dutch lemma log frequency (Baayen et al., 1995) (see Table 4.2 for averages). Assignment of each subset to an interference condition (learned in Spanish vs. not learned in Spanish) was counterbalanced across participants.

Table 4.2
Item characteristics in Experiment 1.

	Interference set (= learned in Spanish)			No interference set (= not learned in Spanish)		
	M	*SD*	*range*	*M*	*SD*	*range*
Spanish word length (in syllables)	2.66	0.69	2-4	2.72	0.70	2-4
English word length (in syllables)	1.39	0.49	1-2	1.40	0.50	1-2
Dutch Celex log frequency	0.94	0.56	0-2.32	0.93	0.55	0-2.32
Dutch Celex per million frequency	22.09	35.68	0-208	20.48	34.37	0-208
Within-set semantic distance[a]	0.81	0.09	0.31-1.10	0.81	0.09	0.39-1.10
	M	*SD*	*range*			
Spanish Levenshtein distance[b]	6.10	1.49	2-10			
English Levenshtein distance[b]	5.18	1.30	2-9			
Across-set semantic distance[a]	0.81	0.10	0.31-1.08			

Note. Item sets differed across participants, as described in the Item selection section. Means (*M*) and standard deviations (*SD*) were first calculated per subject and set, and subsequently averaged over participants. Ranges show the absolute minimal and maximal values per group and set. [a]Semantic similarity was assessed as explained in Chapter 2 (see Appendix A.2). [b]Levensthein distances were calculated between all words in the entire list (regardless of interference set).

4.2.1.3 | *Procedure*

Participants were tested individually in a quiet room adjacent to the experimenter's room. The door between the rooms was kept open at all times for communication and response coding. The experiment consisted of three parts: an English pre-test to determine an item set for the remainder of the experiment, a Spanish learning phase, and finally a surprise English post-test to assess the effect of the Spanish learning phase on the accessibility of the corresponding English words. Participants were led to believe that the study was about learning Spanish. There was no mention of English in the study description other than the fact that participants would need to take an English vocabulary test at the beginning of the study. The post-test thus came as a surprise to participants (as also confirmed in a post-experiment interview). For technical details regarding the set-up and precise stimulus timings for each of the tasks described below, see Appendix C.3.

4.2.1.3.1 | **English Pre-test**

To select a matched participant-specific item set, participants initially completed an English picture-based vocabulary test. They saw 103 pictures of everyday objects, one at a time, and their task was to name them in English to the best of their knowledge. There was no time limit. The experimenter coded participants' answers for correctness via a button press, which immediately initiated the next trial. Participants did not receive feedback on their answers. An item was considered known, and thus suitable for the experiment, only if the participant was able to name the picture correctly on their first try. Synonyms and partial answers were not considered correct. Next to accuracy, naming speed was measured and later used as a baseline for the English post-test.

4.2.1.3.2 | **Spanish Learning Tasks**

The Spanish learning phase started once a participant-specific item set had been created. As in Chapter 2, the learning phase consisted of a mix of receptive and productive tasks, which started out easy and got progressively more difficult. The learning tasks are identical to those in Chapter 2 (for details on stimulus timings see Appendix A.4). The first of those tasks was a familiarization task, in which participants saw and listened to all words and pictures once at their own pace. They were told to repeat the words out loud to start practicing their pronunciation, and were furthermore asked to inform the experimenter if they knew any of the Spanish words already. Spanish words that participants knew already were later excluded from analysis (see section 4.2.1.4).

Subsequently, participants completed two rounds of a two-alternative forced choice task. A picture was presented on screen together with two Spanish words from the 23 to-be-learned words, and participants had to select the label that corresponded to the picture. They received feedback on their choice (a green square around the picture for a correct answer, a red square for an incorrect answer, always accompanied by the correct Spanish word and its audio) and moved on to the next trial. In the second round of this task, participants were asked to try to name the picture before seeing the two Spanish labels. This was done to encourage them to start engaging with the words more actively. This multiple-choice task was followed by two rounds of a word completion task, in which each picture was accompanied by the first syllable of that word (or the first grapheme for monosyllabic words). Participants had to complete the word by saying it out loud. Their answer was then coded by the experimenter for correctness by way of a button press, which initiated the feedback screen (again a green or a red square, the Spanish word and its audio).

Next, participants saw each picture again and were asked to write its Spanish name down on paper and to subsequently press a button to see and hear the correct Spanish word on screen and correct themselves (on paper) based on this feedback. The final learning task was an adaptive picture naming task. Participants first went through two rounds of simply naming all pictures again, during which they received feedback as usual. If after these two rounds there were still words left that they were unable to name properly, the task would continue and those unknown words would be repeated until the participant had named each of them correctly at least twice in a row. These optional, subsequent naming rounds always consisted of at least ten pictures to avoid making the task too easy when only a few words were left to be learned. If necessary, the set of still to-be-learned words was thus complemented by already known items. The task script, however, took care to repeat each of the already known words equally often. The adaptive task continued until either all items had been learned (i.e., been produced correctly twice in a row), or otherwise until 50 minutes of the learning phase had passed. The learning phase then ended in one final round of picture naming without feedback, which was administered to establish which words had been learned successfully in Spanish and which had not.

The presentation order of words was semi-random, such that it was different for every learning task and every participant, but the same across (within-participant) repetitions of a task. This pseudo-random order was chosen to avoid order effects, while at the same time keeping the distance between repetitions within a task constant. The learning phase resulted in a minimum of nine exposures per word with feedback, in addition to the final test without feedback. The number of additional exposures to each item depended on the length of the adaptive picture naming task (as in Chapter 2, see Appendix A.4 for details); on average, participants required 12.96 exposures per item (mean *SD* = 3.18, abs. range = 10-32).

In total, the learning session took a maximum of one hour. To continue with the experiment, participants had to have learned at least 18 of the 23 words. All participants satisfied this criterion (see section 4.2.2.2 for details).

4.2.1.3.3 | **English Post-test**

Finally, participants were tested again in English on all 46 words chosen in the pre-test. Participants saw each picture once. There was no time limit for them to provide their answer and they did not receive feedback. Next to accuracy, naming latencies were measured.

4.2.1.3.4 | **LexTALE**

Finally, participants completed the English version of the LexTALE, a lexical decision based vocabulary test (Lemhöfer & Broersma, 2012).

4.2.1.4 | *Response Coding and Exclusion Criteria*

4.2.1.4.1 | **Naming Accuracy**

Participants' English productions in the post-test were coded as either correct or incorrect/unknown. When participants corrected themselves, or otherwise needed multiple attempts to name a picture, only the last utterance was scored. A word was considered unknown if a participant could not remember the word and said 'I don't know', or if that last utterance was incorrect, but the latter scenario never happened (i.e., all errors were instances of 'I don't know'). Synonyms were not counted as mistakes (0.26% of correct responses were synonyms: 0.4% in the interference condition and 0.2% in the no interference condition). Words that were not successfully learned in Spanish (*M* = 0.77 out of a total of 23, i.e., on average 3%, range = 0-4) and words that were known in Spanish prior to the experiment (*M* = 0.19 out of a total of 23 words, on average 1%, range = 0-2; as determined in the familiarization of the Spanish learning phase), were excluded from analysis.

4.2.1.4.2 | **Naming Latencies**

Naming latencies in the English pre- and post-tests were measured manually in Praat (version 5.3.78, Boersma, 2001) and reflect the time from picture presentation to speech onset. In the analysis, to take initial between-item accessibility differences into account, we calculated difference scores using the English pre-test latencies as the baseline for latencies during the post-test. Because post-test latencies were typically shorter than pre-test latencies, we subtracted post-test latencies from pre-test latencies. The resulting difference scores thus reflect a *speed-up* in naming latencies from pre- to post-test in English. Such a speed-up is to be expected at the very least for the not interfered items, because at post-test in English, participants name the pictures for the second time and hence should be faster in doing so than when they first saw and named the pictures in English (i.e., at pre-test). Crucially, we hypothesized this speed-up to be significantly modulated by Spanish learning and consequently to be smaller for interfered than not interfered items.

Next to the exclusions mentioned above for accuracy, for the naming latency analysis, we additionally excluded trials in which participants were unable to name a picture, named it incorrectly or took multiple attempts at naming. Trials in which participants corrected themselves, coughed or laughed were also excluded. Smacks and hesitations were accepted though; naming latencies for these trials were measured at the onset of the actual word production. These extra exclusions resulted in an average of 3% additional data loss per participant (M_{Int} = 4%, range = 0-17%; M_{NoInt} = 3%, range = 0%-13%). Finally, trials in which participants used an article before the noun in the pre-test, but not the post-test, or vice versa, were also excluded from the latency analysis, because the pre-post naming latency comparison was impossible for these trials. Exclusions due to these article trials resulted in an additional loss of on average 8% of data (M_{Int} = 9%, range = 0-56% M_{NoInt} = 7%, range = 0%-56%). Participants for whom all exclusions taken together resulted in more than 30% data loss in either the interference or the no interference condition were excluded from the naming latency analysis (N = 3). The remaining 23 participants had an average of 41 out of 46 trials left for analysis (range 33-46 trials, M_{Int} = 20.04, M_{NoInt} = 21.30). Note that this means that the naming latency analysis is based on fewer participants than the accuracy analysis.

4.2.1.5 | *Modelling*

We analyzed the data using (generalized) mixed effects models with the lme4 package (version 1.1-21, Bates et al., 2015) in R (Version 3.5.1, R Core Team, 2018). The accuracy data were analyzed using a generalized mixed effects model of the binomial family, fitted by maximum likelihood estimation, using the logit link function and the optimizer 'bobyqa'. The dependent variable consisted of 1's and 0's for correct and incorrect words respectively. The only fixed effects variable was Interference (two levels: no interference, interference) and it was effects coded (-0.5, 0.5). This means that negative beta coefficients reflect more errors for interfered items than not interfered items. Random effects were fitted to the maximum structure justified by the experimental design (Barr et al., 2013), which included random intercepts for both Subjects and Items, as well as random slopes by Subject for Interference. Random slopes that correlated highly ($r > 0.95$) with their respective intercepts were removed to avoid over-fitting. All p-values were calculated by model comparison, omitting one factor at a time, using chi-square tests.

Naming latencies were analyzed using a linear mixed effects model fitted by restricted maximum likelihood estimation (using Satterthwaite approximation to degrees of freedom). Raw naming latencies were first log-transformed and then difference scores (pre-test RT – post-test RT) were calculated and entered into the model. The fixed and random effects structure was identical to the accuracy model. Interference was again effects coded (-0.5, 0.5), meaning that negative beta coefficients in the RT model reflect a smaller latency speed-up from pre- to post-test in English for interfered as compared to not interfered items.

4.2.2 | Results

4.2.2.1 | *English Pre-test Performance*

On average, participants knew 77% (*SD* = 14%, range = 49%–93%) of all 103 words from the English pre-test.

4.2.2.2 | *Spanish Learning Performance*

On average, participants learned 97% (*SD* = 5%, range = 83%–100%) of the 23 Spanish words.

4.2.2.3 | *English Final Test Performance*

4.2.2.3.1 | Naming Accuracy

The mean percentage of correctly recalled English words for the two interference conditions at final test in English is shown in the left panel of Figure 4.1. The model output is shown in Table 4.3. There was no effect of Interference on participants' English post-test production accuracy. Words for which the Spanish translation equivalents had been learned were thus recalled as often as words for which that was not the case.

4.2.2.3.2 | Naming Latencies

The mean naming latency speed-up from pre- to post-test in English for the two interference conditions is plotted in the right panel of Figure 4.1 (raw naming latencies at English pre- and post-test can be inspected in Appendix C.4). Model outcomes can be found in Table 4.3. As expected, we observed a main effect of Interference, such that the difference in naming latencies from pre- to post-test was smaller (i.e., less speeding up) in the interference than in the no interference condition.

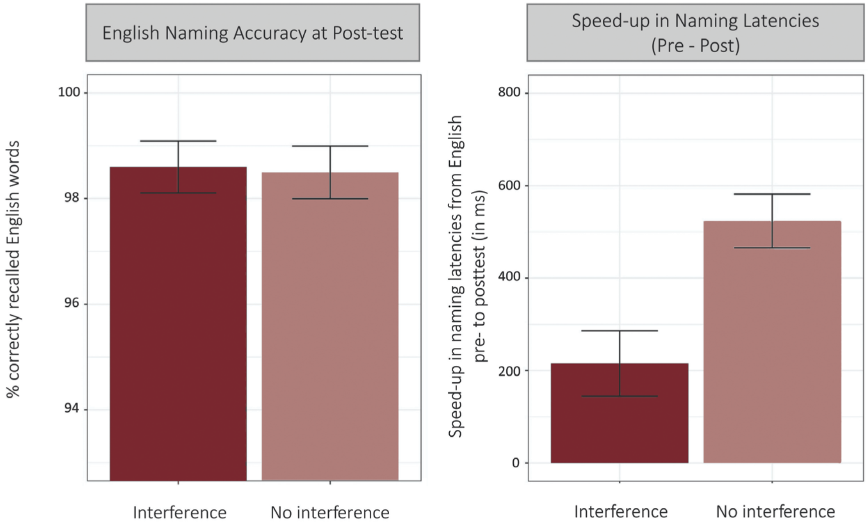

Figure 4.1

Accuracy scores and naming latency speed-up (pre-post; in ms) at final test in English. Error bars reflect the standard error around the condition means.

Table 4.3

Mixed effects model output for naming accuracy and naming latencies in Experiment 1.

	Naming accuracy				Naming latencies			
Fixed effects	Estimate	*SE*	*z*	$p(\chi^2)$	Estimate	*SE*	*t*	$p(\chi^2)$
Intercept	5.58	0.75	7.43	**<.001**	0.22	0.03	7.63	**<.001**
Interference	0.38	1.03	0.37	0.708	-0.24	0.03	-7.28	**<.001**
Random effects	Groups	Var	*SD*	Corr	Groups	Var	*SD*	Corr
Item	Intercept	1.92	1.39		Intercept	0.02	0.15	
Subject	Intercept	0.65	0.80		Intercept	0.01	0.09	
	Interference	4.24	2.06	0.29	Interference	0.00	0.06	0.63

Note. Significant effects are marked in bold. *SE* = standard error; $p(\chi^2)$ = Chi-square p-value; Var = variance; *SD* = standard deviation; Corr = correlation.

4.2.3 | Discussion

In Experiment 1 we asked whether learning a new foreign language comes at the expense of retrieval ease of other, previously learned foreign languages. Participants learned new Spanish words, and we assessed what this new learning did to the accessibility of their (previously known) English translation equivalents in a subsequent English picture naming test. While English words were still recalled correctly at post-test regardless of whether they had been learned in Spanish or not, participants sped-up less from pre- to post-test for words for which they had learned Spanish translations compared to words for which they had not. Although we did not find evidence for interference in recall accuracy, Spanish learning thus hindered the expected latency speed-up in subsequent English productions. Hence, we can conclude that learning words in a new language does come at the cost of at least retrieval *ease* for words in previously learnt foreign languages.

To our knowledge, we are the first to experimentally demonstrate a detrimental effect of learning a new foreign language on a previously learned FL. While previous research had shown that repeatedly retrieving already known translation equivalents hampers subsequent access to just recently learned L3 words (e.g., Bailey & Newman, 2018; Chapter 2), the present study shows that the opposite is also true: new learning negatively affects retrieval ease in a long before acquired foreign language. The present study thus differs from earlier studies in two crucial ways. First, instead of the retrieval of known words, it is the new learning of L3 words that leads to interference. What is more, in the present study, the English translation equivalents were long known to the participants rather than taught to them within the scope of the experiment. The experimental items in the present study were thus no longer episodic memories, but instead 'old', semantic memories, and hence much harder to interfere with than the newly-learned items in, for example, Chapter 2.

Given these differences, it is perhaps not surprising that we did not observe accuracy effects in the present study. In fact, the assumption that 'old' memories are harder to interfere with is also supported by some retroactive interference studies that found that *consolidated* (as compared to unconsolidated) material is less susceptible, if at all, to interference from subsequent learning of new information (e.g., Ellenbogen et al., 2016; Landauer, 1974). In general, it is assumed that the more time a memory has to consolidate, the less it will suffer from subsequent interfering tasks (see Müller & Pilzecker, 1900, for the original formulation of this argument). Whether this is really the case is still being debated (see section 4.4 and Wixted, 2004, for a comprehensive discussion of the role of consolidation in RI). Regardless though, the idea that consolidated memories are *resistant* to interference is at odds with the fact that we do find clear evidence for interference of Spanish word learning on English words in reaction times. Reaction times have often been ignored in the RI literature, despite their potential to uncover nuances of retrieval difficulty that would go unnoticed with a dichotomous 'remembered – not remembered' response coding (see Postman & Kaplan, 1947, for a more elaborate account of this argument). Clearly, new learning can lead to retrieval difficulties of well consolidated material, just possibly not to the extreme extent (i.e., complete retrieval failure) that is usually probed in RI studies.

Outside of the RI literature, a slow-down in retrieval speed has been interpreted as evidence for the early stages of forgetting (see Chapter 2). Yet, we clearly did not succeed at truly making our participants *forget* English by teaching them Spanish. Are such extreme interference effects impossible to model in the lab, or do they simply take longer to set in? Even in real life, it is very possible that the detrimental effects of learning a new language do not surface until a few months into the learning process. The newly learned Spanish words, being fragile and not yet consolidated themselves, might not yet interact with translation equivalents from other languages in the way necessary for maximal interference effects to arise.

Between-language interference effects in production are usually explained in terms of lexical competition between translation equivalents (e.g., Chapter 2). From research on novel word learning, however, it appears that some types of lexical competition do not emerge immediately, but instead require consolidation. Dumay and Gaskell (2007), for example, showed that during word recognition newly learned words only compete with phonological neighbours after a night of sleep (see also Bakker et al., 2014; Bowers et al., 2005). Similarly, novel words have been shown to become increasingly word-like in their online processing signatures only after an offline consolidation period including a night of sleep (Bakker et al., 2015a, 2015b). While these studies investigate competition effects between form-similar words in perception rather than competition between translation equivalents in production, it is possible that the same principles hold for between-translation competition and hence that newly acquired vocabulary also needs to consolidate first before it can compete with translation equivalents. Davis and Gaskell (2009) explain their findings (with form-similar word pairs) by assuming that novel words, just like any new memories, are initially encoded as episodic traces that are heavily dependent on the hippocampus. Through offline consolidation, and particularly through sleep, these episodic traces become gradually integrated into the existing neocortical memory network, and hence only then compete with related words, such as phonological neighbours, in the mental lexicon. Research suggests that this consolidation process is aided by the first night of sleep (Dumay & Gaskell, 2007), but also that it is a gradual process that can take up to multiple weeks to complete (Takashima et al., 2006).

If novel words require such a slow integration process, it might seem puzzling that we found any interference effects at all, even in reaction times. There is, however, also evidence for immediate lexical integration (e.g., Borovsky et al., 2012; Coutanche & Thompson-Schill, 2014; Lindsay & Gaskell, 2013). Interestingly, outside of the lexical domain, integration has been shown to be especially fast when the newly acquired knowledge fits in with existing knowledge (e.g., van Kesteren et al., 2010; Zhang et al., 2018). Since newly learned translation equivalents share their concept with existing words, it is very possible that they can benefit from such fast integration. The RT effect, in our eyes, then reflects the immediate beginning of the lexical integration process, allowing the Spanish words to interact with English words and causing a slow-down in subsequent retrieval of the latter. Possibly, however, they were not yet integrated enough to entirely block access to their English translation equivalents (as would be apparent in an accuracy effect). Assuming that consolidation is a gradual process that is crucially aided by sleep, the newly learned Spanish words might become much stronger interferers after a longer consolidation time window including a night of sleep. Experiment 2 addresses this possibility by separating the final English test from the Spanish learning by roughly 24 hours.

In addition to manipulating the consolidation time window for the newly learned Spanish words, we also changed the timing of the pre-test. The reason for this change relates to an alternative explanation for the RT effect. Research has shown that despite the initial consolidation process that makes memories less prone to interference, memories do not necessarily stay stable forever. Walker et al. (2003) showed that retrieval of a consolidated memory in fact can make it labile and hence susceptible to interference again. The destabilized memory then needs to be 're-consolidated', a process that is faster than the original consolidation, but that can still take up to six hours (Stickgold & Walker, 2005). If retrieving a memory destabilizes it, we might have artificially increased our chances of interfering with the English words by having participants retrieve them in the pre-test, immediately before some of them were learned in Spanish. To avoid this potential confound and to assess the robustness of our interference effect, we separated the English pre-test from the Spanish learning phase by roughly 24 hours (a time frame that is long enough to allow for complete reconsolidation, should it be necessary, see Stickgold & Walker, 2005).

To take both the pre- and post-test timing changes into account, we needed to test two groups of participants in Experiment 2. Both groups differed from the group tested in Experiment 1 in that they completed the English pre-test one day before the Spanish learning took place. The two groups in Experiment 2, however, differed from each other in the timing of the English post-test: the no-consolidation group was tested in English immediately after the Spanish learning phase (i.e., as in Experiment 1), while the consolidation group was sent home to consolidate the newly learned Spanish words and was tested in English a day later. We hypothesized that interference effects would be stronger, as apparent in an effect in accuracy and/or a bigger naming latency effect, for the group that gets time to consolidate the Spanish words compared to the group that does not. Furthermore, a comparison of the no-consolidation group with the group tested in Experiment 1 will address whether the RT effects found in Experiment 1 are reliable and still obtained even when an interruption of reconsolidation can no longer be the reason for interference success.

4.3 | Experiment 2

The experimental set-up of Experiment 2 was very similar to that of Experiment 1. Below we report only methodological differences between the two.

4.3.1 | Method

4.3.1.1 | *Participants*

A total of 86 Dutch native speakers with normal or corrected-to-normal vision and without a history of neurological or language-related impairments participated in Experiment 2. None of them had taken part in Experiment 1. Three of them had to be excluded from the analysis either because they did not learn enough Spanish words (N = 1), or because they failed to show up for the final experimental session (N = 2). One additional participant had to be excluded due to a technical failure (N = 1). The remaining 82 participants (57 female) were between 18 and 29 years of age (*M* = 21.99) and had Dutch as their only mother tongue. All but four of the participants indicated having no prior knowledge of Spanish. The four participants who did report having learned some Spanish in the past, had done so for very short amounts of time (*M* = 4.75 months, range = 1-9), rated their current knowledge of Spanish as very poor (*M* = 1.38; *SD* = 0.48, abs. range = 1-2, on a scale from 1-7) and stated that they hardly ever spoke Spanish. Despite the prior exposure, their knowledge of Spanish can thus be described as minimal. Furthermore, as in Experiment 1, all participants reported English as their first and most frequently spoken foreign language (see Table 4.4 for reports of frequency of use, proficiency self-ratings and English LexTALE scores). Other foreign languages participants knew again included most prominently French, German and Latin.

Upon coming to the lab, participants were randomly assigned to either the consolidation group (N = 41) or the no-consolidation group (N = 41). As confirmed by independent t-tests, the two groups did not differ in their average frequency of use or their performance on the English LexTALE and were also comparable on their English proficiency self-ratings (see Table 4.4).

Table 4.4
Participant characteristics - Experiment 2.

	Consolidation group (N=41)			No-consolidation group (N=41)		
	M	*SD*	*range*	*M*	*SD*	*range*
English AoA	10.56	1.61	6-14	10.90	1.56	7-14
English LoE (years)	9.54	4.24	2-22	10.20	2.84	4-15
English Frequency of Use (min/day)						
Speaking	17	24	0-127	21	53	0-240
Listening	145	73	10-300	149	113	0-600
Reading	85	65	0-240	83	79	0-300
Writing	29	43	0-240	28	57	0-300
English Proficiency[a]						
Speaking	5.27	1.00	3-7	4.98	1.19	1-7
Listening	6.10	0.77	4-7	5.68	1.08	2-7
Reading	5.90	0.89	3-7	5.83	0.86	3-7
Writing	5.15	1.04	2-7	4.71	1.25	1-7
English LexTALE	74.3	12.03	51-95	72.9	11.00	48-92

Note. *M* = mean; *SD* = standard deviation; AoA = Age of acquisition; LoE = length of exposure (i.e., amount of years participants had been learning English). [a]Proficiency self-ratings were given on a scale from 1(very poor) – 7(like a native speaker).

4.3.1.2 | *Materials*

The stimulus database from Experiment 1 was extended to 140 concrete and non-cognate words referring to everyday objects or animals (the full list can be inspected in Appendix C.5). We did so to ensure that the item selection script (same as in Experiment 1) would succeed at constructing a participant-specific item list in as many cases as possible, and hence to avoid large drop-out rates at pre-test (as noted above, four participants had to be excluded from Experiment 1 because of poor performance on a list of 103 words). The words were between one and five syllables long in Spanish (*M* = 2.56, *SD* = 0.68), and between one and four syllables long in English (*M* = 1.42, *SD* = 0.60). The Celex frequencies of the corresponding Dutch translation equivalents ranged from 0 to 818 occurrences per million (*M* = 29.02, *SD* = 76.69, Dutch lemma frequencies, Baayen et al., 1995). For matching items in the interference and no interference sets, we again used the corresponding log frequencies, which ranged from 0 to 2.91 ($M = 0.98$, $SD = 0.65$). Pictures were identical to those in Experiment 1, with additional pictures taken from Google images. Audios were recorded by the same female Spanish native speaker from Andalucía (Spain) as in Experiment 1.

4.3.1.2.1 | **Item Selection**

As in Experiment 1, a Matlab script created a participant-specific item set of 46 known English words based on each participant's pre-test performance. The item selection and replacement procedure was identical to that in Experiment 1 (mean words replaced = 17.81; range = 3-36, mean of 16.12 in consolidation group, mean of 19.49 in no-consolidation group, see Appendix C.2). As can be seen in Table 4.5, words in the two interference subsets were again matched on Spanish and English word length, within- and across-set semantic similarity, as well as on phonological similarity in English and Spanish and Dutch frequency.

Table 4.5
Characteristics of items used in Experiment 2.

	Consolidation group						No-consolidation group					
	Interference set			No interference set			Interference set			No interference set		
	M	*SD*	*range*	*M*	*SD*	*range*	*M*	*SD*	*range*	*M*	*SD*	*range*
Spanish word length (in syllables)	2.60	0.70	2-5	2.62	0.69	2-5	2.60	0.67	2-5	2.60	0.69	2-5
English word length (in syllables)	1.36	0.50	1-4	1.39	0.50	1-3	1.40	0.53	1-4	1.37	0.50	1-3
Dutch log frequency	1.02	0.58	0-2.32	1.02	0.59	0-2.91	1.05	0.58	0-2.91	1.05	0.59	0-2.32
Dutch frequency per million	22.60	31.76	0-208	23.44	35.98	0-818	24.90	35.47	0-818	25.00	34.54	0-208
Within-set semantic distance[a]	0.83	0.09	0.31-1.09	0.83	0.11	0-1.09	0.82	0.09	0.37-1.09	0.81	0.10	0-1.09
	M	*SD*	*range*				*M*	*SD*	*range*			
Spanish Levenshtein distance[b]	5.94	1.52	2-10				5.94	1.52	2-10			
English Levenshtein distance[b]	5.05	1.23	2-11				5.08	1.23	2-11			
Across-set semantic distance[a]	0.83	0.08	0.37-1.11				0.83	0.08	0.40-1.11			

Note. Item sets differed across participants, as described in the Item selection section. Means (*M*) and standard deviations (*SD*) were first calculated per subject and interference condition, and subsequently averaged over groups. Ranges show the absolute min and max values per group and condition. [a]Semantic similarity was assessed as explained in Chapter 2 (see Appendix A.2). [b]Levensthein distances were calculated between all words in the entire list (regardless of interference set).

4.3.1.3 | *Procedure*

Unlike Experiment 1, Experiment 2 took place over multiple days (see Figure 4.2 for a schematic representation of the experimental set up for both experiments). On day 1, participants came to the lab to take the English pre-test. One day later (i.e., day 2), they returned for the Spanish learning session (time between the two sessions: consolidation group: M = 23.46 hours, SD = 2.23, range = 19-28.5; no-consolidation group: M = 23.29 hours, SD = 2.82, range = 16-29). On the same day, the no-consolidation group also completed the final English post-test. The consolidation group instead was sent home after the Spanish learning phase and only took the final English post-test on day 3, after another night of sleep (after M = 24.12 hours, SD = 2.75, range = 18.5-30). All tasks were administered exactly as in Experiment 1. The only difference in tasks between the two experiments concerns the final Spanish test: in Experiment 2, participants underwent two final Spanish tests without feedback (as opposed to just one in Experiment 1). The consolidation group did the first of those post-tests on day 2 immediately after learning (i.e., as in Experiment 1) and the second on day 3 before the final English post-test. The no-consolidation group, in turn, did both Spanish post-tests in immediate succession after the Spanish learning phase on day 2. The second Spanish test was added primarily to assess whether the participants in the consolidation group had forgotten any words overnight. Note that next to providing a measure of overnight Spanish retention for the consolidation group, the additional Spanish test on day 3 also served to match the two groups in terms of recency of exposure to Spanish prior to the final English post-test, such that the only difference between the two groups was ultimately whether or not the Spanish words got time to consolidate overnight.

4.3.1.4 | *Coding and Exclusion Criteria*

4.3.1.4.1 | Naming Accuracy

Participants' answers were scored as in Experiment 1. Only trials in which participants were either entirely unable to name the picture (consolidation group: 83% of all errors, no-consolidation group: 92% of all errors), or named it incorrectly (consolidation group: 17% of all errors, no-consolidation group: 8% of all errors) were counted as errors, synonyms were not (consolidation group: 0.5% of all correct answers, no-consolidation group: 0.8% of all correct answers). Moreover, words which participants already knew in Spanish before starting the experiment ($M_{\text{ConsolGroup}}$ = 0.20 out of 23, 1%, $\text{range}_{\text{ConsolGroup}}$ = 0-2, $M_{\text{NoConsolGroup}}$ = 0.27 out of 23, 1%, $\text{range}_{\text{NoConsolGroup}}$ = 0-2), words that were not successfully learned in Spanish, as assessed during the second Spanish post-test ($M_{\text{ConsolGroup}}$ = 1.85 out of 23, 8%, $\text{range}_{\text{ConsolGroup}}$ = 0-7, $M_{\text{NoConsolGroup}}$ = 1.24 out of 23, 5%, $\text{range}_{\text{NoConsolGroup}}$ = 0-7) as well as words that had accidentally been coded as correct in the pre-test, but that were actually unknown to participants in English ($M_{ConsolGroup}$ = 0.12 out of 23, 1%, $range_{ConsolGroup}$ = 0-1, $M_{NoConsolGroup}$ = 0.12 out of 23, 1%, $range_{NoConsolGroup}$ = 0-3) were again excluded from all subsequent analyses.

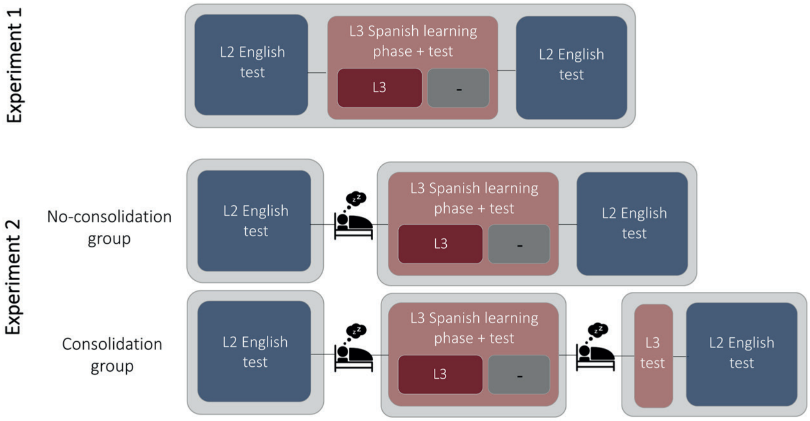

Figure 4.2
Schematic representation of the experimental set-up for both Experiment 1 and 2. Grey boxes in the background indicate separate testing days.

4.3.1.4.2 | Naming Latencies

As in Experiment 1, trials excluded from the accuracy analysis were also excluded from the RT analysis. On top of that, as in Experiment 1, trials in which participants made errors, took multiple attempts at naming, corrected themselves or coughed or laughed were excluded from RT analysis (6% of trials on average, $M_{ConsolGroup}$ = 5%, M_{ConInt} = 7%, $M_{ConNoInt}$ = 4%; $M_{NoConsolGroup}$ = 6%, $M_{NoConInt}$ = 8%, $M_{NoConNoInt}$ = 5%). Additionally, trials in which participants inconsistently used articles at pre- but not post-test in English (or vice versa) were also excluded from RT analysis, resulting in an additional loss of on average 3% of trials per person ($M_{ConsolGroup}$ = 2%, M_{ConInt} = 2%, $M_{ConNoInt}$ = 3%; $M_{NoConsolGroup}$ = 3%, $M_{NoConInt}$ = 3%, $M_{NoConNoInt}$ = 3%). As in Experiment 1, participants who after all exclusions had less than 70% of trials in either of the two interference conditions left, were excluded from RT analysis ($N_{ConsolGroup}$ = 3, $N_{NoConsolGroup}$ = 1). The remaining 78 participants had an average of 41 of 46 trials left ($M_{ConsolGroup}$ = 41.34, range = 37-46, M_{ConInt} = 19.22, $M_{ConNoInt}$ = 21.46; $M_{NoConsolGroup}$ = 40.53, range = 32-46, $M_{NoConInt}$ = 19.24, $M_{NoConNoInt}$ = 21.07).

4.3.1.5 | *Modelling*

As in Experiment 1, the data were analyzed using (generalized) mixed effects models with lme4 in R. Fixed effects were Interference (two levels: no interference,

interference) and Group (two levels: consolidation, no-consolidation) as well as their interaction. Both fixed factors were effects coded (-0.5,0.5). All other model specifications were as in Experiment 1.

4.3.2 | Results

4.3.2.1 | *English Pre-Test Performance*

Participants' performance on the English pre-test did not differ significantly between groups. Participants knew on average 67% (consolidation group, *SD* = 12, range = 41%-89%) and 62% (no-consolidation group, *SD* = 14, range = 34%-83%) of the words from the English pre-test ($t(78.11) = 1.75$, $p = .085$).

4.3.2.2 | *Spanish Learning Performance*

Participants in both groups were successful at learning the 23 Spanish words and did not differ statistically speaking in their learning performance, both as assessed in the first ($t(69.74) = 1.50$, $p = .138$, $M_{\text{ConsolGroup}}$ = 97%, *SD* = 4%, range = 83%-100%; $M_{\text{NoConsolGroup}}$ = 95%, *SD* = 6%, range = 65%-100%) and the second Spanish post-test ($t(76.26) = -1.69$, $p = .10$, $M_{\text{ConsolGroup}}$ = 92%, *SD* = 8%, range = 70%-100%; $M_{\text{NoConsolGroup}}$ = 95%, *SD* = 6%, range = 70%-100%). For the no-consolidation group, there was no significant difference in performance between the two post-tests, as expected ($t(40) = 0.54$, $p = .59$). For the consolidation group, however, there was a difference, with participants making more errors on the second post-test in Spanish than on the first, showing that on average participants did forget some of the words overnight ($t(40) = 4.49$, $p < .001$).

4.3.2.3 | *English Final Test Performance*

4.3.2.3.1 | **Naming Accuracy**

Average naming accuracy scores per group and interference condition can be found in Figure 4.3 and model outcomes for accuracy scores can be found in Table 4.6. Unlike in Experiment 1, we observed a main effect of Interference on recall accuracy, such that interfered items were recalled less well than not interfered items. There was no effect of Group, nor was the interaction between Interference and Group significant.

4.3.2.3.2 | **Naming Latencies**

Mean latency speed-up from pre- to post-test in English for both groups and interference conditions is shown in Figure 4.4, and model outcomes for naming latencies are reported in Table 4.6 (for raw naming latencies see Appendix C.4). As in Experiment

1, we observed a main effect of Interference, indicating that participants sped-up more from pre- to post-test in the no interference condition than the interference condition. Learning words in Spanish thus partially impeded (i.e., reduced) the expected latency speed-up in English productions from pre- to post-test (remember that naming at pre-test is more difficult than at post-test because the pre-test is the first time around that participants see the pictures and retrieve the English labels). There was no main effect of Group and no interaction between the two fixed factors.

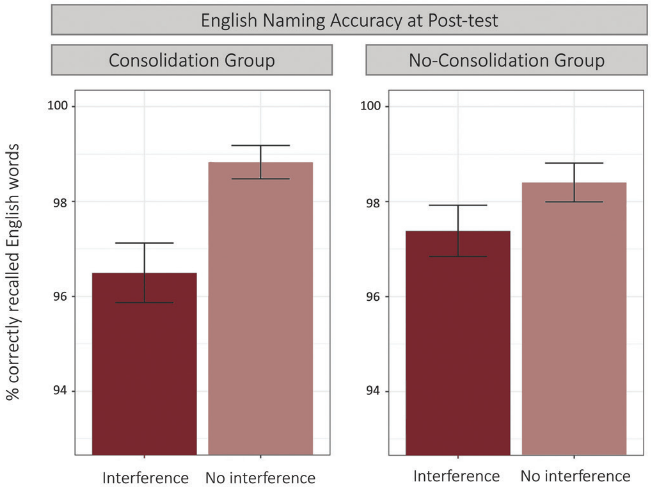

Figure 4.3

Accuracy scores at final test in English in Experiment 2. Error bars reflect the standard error around the condition means.

Table 4.6

Mixed effects model output for naming accuracy and naming latencies in Experiment 2.

	Naming accuracy				Naming latencies			
Fixed effects	Estimate	*SE*	*z*	$p(\chi^2)$	Estimate	*SE*	*t*	$p(\chi^2)$
Intercept	4.30	0.24	17.73	**<.001**	0.23	0.03	8.82	**<.001**
Interference	-0.80	0.25	-3.23	**.001**	-0.17	0.02	-7.42	**<.001**
Group	0.05	0.26	0.18	.856	0.01	0.04	0.22	.821
Interference * Group	-0.58	0.50	-1.17	.241	0.01	0.05	0.09	.926
Random effects	Groups	Var	*SD*	Corr	Groups	Var	*SD*	Corr
Item	Intercept	1.20	1.09		Intercept	0.02	0.15	
Subject	Intercept	0.08	0.28		Intercept	0.03	0.17	
					Interference	0.02	0.12	0.54

Note. Significant effects are marked in bold. *SE* = standard error; $p(\chi^2)$ = Chi-square p-value; Var = variance; *SD* = standard deviation; Corr = correlation.

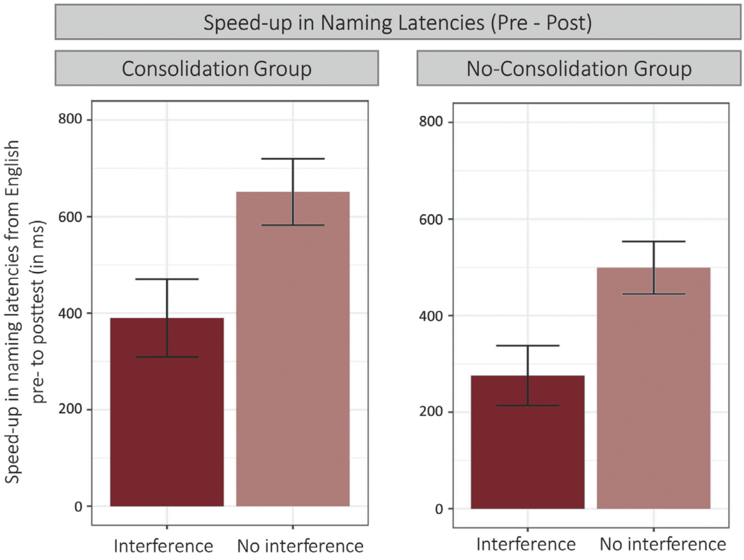

Figure 4.4

Naming latency speed-up (pre-post; in ms) in English productions in Experiment 2. Error bars reflect the standard error around the condition means.

4.3.2.4 | *Joint Analysis of Experiments 1 and 2*

To test for differences in interference magnitude between the group in Experiment 1 and the two groups in Experiment 2, we ran mixed effects models for both accuracy and naming latencies with Group as a 3-level factor (repetition contrast coded, first contrast compares the no-consolidation group with the group from Exp.1, and the second contrast compares the consolidation group with the group from Exp.1), and Interference (no interference, interference) effects coded (-0.5, 0.5) as usual. The model outcomes can be inspected in Table 4.7. Both for naming accuracy and naming latencies, we found a significant main effect of Interference, indicating that overall participants made more errors and sped-up less from English pre- to post-test in naming pictures for which they had learned Spanish translation equivalents. We observed no main effects of Group in either model, nor did any of the groups differ in the magnitude of their interference effect in naming latencies. In naming accuracy, the consolidation group had a numerically bigger interference effect than the group in Experiment 1; the interaction, however, did not reach statistical significance. The no-consolidation group did not differ from the group in Experiment 1.

Finally, to check whether the apparent difference in main effects in accuracy between experiments is meaningful, we compared the interference effect from Experiment

1 with the combined interference effect (i.e., the overall main effect in accuracy) obtained in Experiment 2. More specifically, we ran a model with a contrast coding that compares the group from Experiment 1 to the average of both groups from Experiment 2 (Helmert contrast) and found no significant interaction ($\beta = 1.14$, $z = 1.55$, $p(\chi^2) = .122$), meaning that the numerical difference between experiments is not substantial and very possibly just due to chance.

Table 4.7
Mixed effects model output for naming accuracy and naming latencies for both experiments combined.

	Naming accuracy				Naming latencies			
Fixed effects	**Estimate**	***SE***	***z***	***p*(χ^2)**	**Estimate**	***SE***	***t***	***p*(χ^2)**
Intercept	4.41	0.22	20.20	**<.001**	0.22	0.02	10.18	**<.001**
Interference	-0.52	0.23	-2.21	**.028**	-0.19	0.02	-9.47	**<.001**
GroupC1: NoConsol-vs-Exp1	0.50	0.34	-1.50	.127	-0.01	0.05	-0.18	.856
GroupC2: Consol-vs-Exp1	-0.46	0.34	-1.34	.178	0.02	0.05	0.36	.711
Interference*GroupC1	0.57	0.60	0.94	.345	-0.06	0.05	-1.12	.255
Interference*GroupC2	-1.15	0.61	-1.88	.060	0.07	0.05	1.26	.202
Random effects	**Groups**	**Var**	***SD***		**Groups**	**Var**	***SD***	**Corr**
Item	Intercept	1.05	1.03		Intercept	0.02	0.14	
Subject	Intercept	0.16	0.40		Intercept	0.02	0.16	
					Interference	0.01	0.11	0.54

Note. Significant effects are marked in bold. *SE* = standard error; $p(\chi^2)$ = Chi-square p-value; Var = variance; *SD* = standard deviation; Corr = correlation.

4.3.3 | Discussion

Experiment 2 was designed to answer two questions. First and foremost, we wanted to know whether allowing the newly learned Spanish words time to consolidate would make them stronger interferers with English translation equivalents. Second, we were interested in whether the interference effect in naming latencies observed in Experiment 1 would persist and replicate even if we separated the English pre-test from the Spanish learning phase (allowing for full reconsolidation unlike in Experiment 1, where the lack thereof may have favored the emergence of interference effects). To that end, we compared two groups of participants, both of whom took the English pre-test a day before learning Spanish, but who crucially differed in their timing of the English post-test: one group had time to consolidate the newly learned Spanish words and was only tested in English after a night of sleep; the other group in turn was tested on English immediately after the Spanish learning phase, without time to consolidate the Spanish words. Results showed an overall interference effect in naming latencies, thus replicating Experiment 1, despite the separation of the English pre-test from the Spanish learning phase. Interestingly, unlike Experiment 1, we also observed an overall interference effect in naming accuracy. Learning Spanish translations thus actually made participants forget some of the corresponding English words (at least temporarily). Neither of those effects, however, were substantially modulated by consolidation time. Based on the present study, we conclude that consolidation, at least after one night of sleep, does not seem to make newly learned words stronger interferers.

4.4 | General Discussion

In this chapter, we asked whether learning a new language can make you forget previously learned foreign languages, and whether such detrimental effects set in immediately after learning, or only later in the learning process. In Experiment 1, we showed that new word learning comes at the cost of subsequent retrieval ease of words in another foreign language: while participants were still able to recall English words after learning their Spanish translation equivalents, the expected speed-up in naming latencies from pre- to post-test was much smaller for those words compared to words for which they had not learned Spanish translations. The goal of Experiment 2 was two-fold. First, we aimed at replicating the result from Experiment 1 while removing the English pre-test from the Spanish learning by a day to eliminate the possibility of artificially facilitating the emergence of interference effects. Second, we asked whether interference effects would grow stronger (and hence also be found in accuracy) if we allowed the newly learned words time to consolidate overnight.

We indeed replicated the naming speed effect from Experiment 1, thus reinforcing the finding that learning words from a new foreign language is detrimental for subsequent retrieval ease of words in other foreign languages. Interestingly, unlike in Experiment 1, we also report a significant interference effect in naming accuracy in Experiment 2, suggesting that learning a new language can under some circumstances actually make you forget corresponding words in another foreign language. These interference effects, however, were *not* modulated by consolidation, and the difference in main effects between experiments, in fact, proved insignificant as well. A statistical model with both experiments together shows overall main effects of interference in accuracy and naming latencies in the absence of significant interactions between groups. Hence, we conclude that learning a new language does hamper subsequent retrieval ease and ability in other foreign languages. Moreover, it appears that these effects can be observed immediately after learning, and do not need time (or overnight consolidation) to emerge.

4.4.1 | Between-Language Interference as a Cause of Attrition

In general, our findings fit in well with previous research on interference-based forgetting: engaging with a foreign language affects your later ability to retrieve words in another foreign language. However, the present study also crucially extends the literature on foreign language attrition. First and foremost, we provide evidence for interference not from the use of other foreign languages, but instead from the *new learning* of foreign language material. In doing so, we provide the first empirical evidence for the anecdotal reports mentioned in the Introduction (section 4.1), namely that learning a new language negatively affects the subsequent retrieval of previously learned foreign languages.

Second, the present experiments demonstrate new learning-induced retrieval difficulty for *well-consolidated* foreign language vocabulary, that is vocabulary learned long before rather than during the experiment. Showing interference effects for old, semantic memory content is arguably more difficult and hence more convincing proof of the role of between-language competition in language forgetting than comparable effects on just recently acquired vocabulary (e.g., Bailey & Newman, 2018; Chapter 2). English is a language that our participants started learning a long time ago and which they have achieved considerable proficiency in. It is thus quite remarkable that a Spanish learning session of just one hour on average made some participants unable to retrieve some of the corresponding English translations. Admittedly, the accuracy effect is numerically very small: though statistically robust, it amounts to a difference between interference conditions of on average less than one word (out of 23). It would be interesting for future research to test whether a longer learning session, spread out over multiple days, allowing not only for consolidation but also for repeated rehearsal of the new language, would show (numerically) stronger interference effects in the lab. Such an experimental set-up would also resemble real life learning situations a little more closely.

The combination of the above mentioned two design aspects (inducing interference through *new learning* and on L2 vocabulary *learned long ago*) is what makes the present experiments novel and different from previous research on language forgetting. Chapters 2 and 3 and Bailey and Newman (2018) both showed forgetting effects on recently learned foreign language words induced via retrieval practice of translation equivalents. Next to those papers, two other studies have addressed similar questions. Isurin and McDonald (2001) were also interested in what new learning does to a previously learned foreign language. They had participants first learn a list of English-Hebrew translation pairs, followed by a list of English-Russian word pairs (or vice versa). While they did find evidence for interference of learning the second list on subsequent retrieval success for the first learned list, note that the two languages in their experiment were learned in immediate succession. Their results thus demonstrate retroactive interference from new learning on still episodic memories, and hence are likely to be the consequence of the interruption of the consolidation process of the first learned list. Levy et al. (2007), in turn, showed that retrieval of L2 English words impairs subsequent recall of L1 translation equivalents. Words in the participants' mother tongue were – like the English words in our study – known long before the experimental session. However, Levy and colleagues probed their participants' memory of L1 words in a rather indirect way (via a rhyme generation and semantic relatedness generation task) which might have underestimated participants' actual L1 knowledge. Moreover, critically, their study differs from ours in that their interference phase consisted of retrieval of known words rather than new learning.

4.4.2 | Retroactive Interference and Foreign Language Attrition

Apart from informing research in the language domain, our study also adds to the memory literature on retroactive interference effects (RI). As mentioned in the discussion of Experiment 1 (section 4.2.3), it is not common for traditional RI studies to report naming latencies. With a few exceptions (e.g., Sanders et al., 1974) and despite an early call to report reaction times next to accuracy statistics (Postman & Kaplan, 1947), the vast majority of studies on the topic to date rely on measures of recall accuracy. Quite in contrast to those traditional memory studies, most robust evidence for RI in our experiments comes from naming latency effects. Although these effects might seem less convincing than effects on retrieval ability, we think they provide just as much evidence for interference processes. Especially in a design where participants do not have a response time limit, naming latency differences between items can capture subtle differences in interference susceptibility that a dichotomous 'remembered – not remembered' accuracy measure does not. This might be particularly true and important when looking at old, well consolidated semantic knowledge, as compared to the episodic memory traces that are typically targeted in RI studies. Without measuring response latencies in Experiment 1, for example, we would have erroneously concluded that learning a new language does not impact previously learned languages. RTs clearly provide a more nuanced picture of retroactive interference effects.

Another difference between our study and *classical* RI studies is that we test for interference on well consolidated semantic knowledge. Remember that in a classical RI experiment, learning of list 1 is immediately followed by learning of list 2, which in turn is followed by an immediate test of list 1 (Barnes & Underwood, 1959). The typical finding is that learning list 2 hampers subsequent retrieval of items on list 1 at the final test (compared to a condition in which no second list is learned), and it is assumed that this happens because list 2 learning interferes with the consolidation of list 1 material, essentially overwriting the latter (Müller & Pilzecker, 1900). Explaining our interference effects through a lack of consolidation of the English words (list 1 items) makes little sense. English words were learned long ago, and so learning Spanish (list 2) should have no effect if retroactive interference is only possible through an interruption of the original consolidation process of list 1 (i.e., English) items. In light of these considerations, the interference effects that we report are surprising. They clearly challenge the assumption that consolidation makes memories resistant to interference (Müller & Pilzecker, 1900) and are at odds with studies that have failed to find retroactive interference effects on consolidated material (e.g., Ellenbogen et al., 2006; Sheth et al., 2012).

Interestingly though, we are not the only ones to report RI-like effects with consolidated material (e.g., Houston, 1967; McGeoch & Nolen, 1933; and see Bailes et al., 2020; Pöhlchen et al., 2020 for failed replication attempts of Ellenbogen et al., 2006). In trying to explain the diverging study outcomes, Wixted (2004) explains that the interference effects that are documented for consolidated (list 1) material are caused by an entirely different mechanism than the classical RI effects on unconsolidated material: interference effects in studies allowing for consolidation of list 1 are, according to Wixted, caused by competition between items on the two lists during retrieval at final test of list 1. Items on both lists are associated with the same cues, and given that items on list 2 have been retrieved more recently as exemplars for those cues, they have a competition advantage over list 1 items at final test and (at least partially) block access to the latter. In studies that involve consolidated list 1 material, it is hence no longer the failure to consolidate that causes retrieval difficulties at final test, but instead competition from the more recent retrieval of competitors associated to the same cue (i.e., list 2 items) during retrieval.

Wixted's (2004) interpretation of RI on consolidated material is in fact very similar to how we (in Chapter 2) as well as other lab-based language attrition studies explain their findings (Bailey & Newman, 2018; Levy et al., 2007). Just like items in the two lists in RI studies are connected to one another via a shared cue, so are the English and Spanish words connected to one another via their shared concept. At final test, the Spanish words, having been retrieved more recently for a given concept, have a competition advantage over their English competitors and hence make it harder for participants to recall the English words, compared to English words for which no Spanish translations were learned. In a nutshell, while it is true that consolidated knowledge, such as the English words in our study, is more difficult to interfere with than unconsolidated material, they are not resistant to interference. Instead, they are subject to different interference dynamics than unconsolidated memories, namely to interference through competition during retrieval.

Finally, how come that some RI studies succeeded at finding RI for consolidated material while others failed? A closer look at these studies' experimental design reveals that they often differ in their timing of the final test (i.e., list 1 test, the English post-test in our experiments). As Wixted (2004) points out, in studies that find RI effects for consolidated list 1 material, the final test of the first learned list always immediately follows the learning of the second, interfering list (e.g., Houston, 1967; McGeoch & Nolen, 1933). When testing of the first list is in turn separated in time from list 2 learning, smaller or no RI effects tend to be observed on consolidated list 1 material (e.g., Postman & Alper, 1946; Sisson, 1939). According to Wixted, the difference in RI magnitude in these studies is due to the temporary nature of the competition advantage that list 2 items have over list 1 items at final test. It appears to wear out with time, attesting for the decrease or the entire absence of RI effects when the final test of list 1 is removed sufficiently in time from list 2 learning.

While, as pointed out earlier, our findings fit in well with the general description of RI on consolidated material, it should be noted that our study shows no evidence for such a weakening of the interference effect as the final test (in English) is removed in time from Spanish learning (i.e., list 2 learning in RI terms; resembling the experimental set-up for the consolidation group in Experiment 2, compared to the no-consolidation group). In fact, as we will discuss more in the next section, the trend in our data goes in the opposite direction, towards a stronger interference effect for the consolidation group compared to the no-consolidation group. It is possible that the stimulus materials used in the present study, that is vocabulary lists that are meaningful also outside of the context of the experiment and that have been learned and consolidated a long time ago[10], adhere to slightly different principles than the rather artificial materials tested in memory studies on retroactive interference. Future studies with similar linguistic material will be important to understand in how far the effects reported in our study are driven by the same or different mechanisms as RI effects in the memory literature. Nonetheless, while our study does not support the role of the length of the time period between interference (list 2 / Spanish learning) and final test (of list 1 / English), our results are in line with the general finding that retroactive interference on consolidated material is possible.

4.4.3 | The Role of Consolidation in Lexical Competition

Assuming that the interference effects that we observed are based on competition between translation equivalents in different languages, and knowing that some competition effects (specifically, lexical competition in spoken word recognition) have been shown to depend on the integration of newly learned words into the existing mental lexicon (Davis & Gaskell, 2009), one might expect interference effects to increase after the newly learned Spanish words have had the chance to consolidate. We addressed this possibility in Experiment 2, yet were unable to confirm this hypothesis. The interference effect in accuracy was numerically bigger in the consolidation group and was in fact only statistically reliable in that group (as tested in post-hoc models for each group separately; consolidation group: β = -1.02, z = -2.63, $p(\chi^2)$ = .008; no-consolidation group: β = -0.53, z = -1.64, $p(\chi^2)$ = .101), but the difference in interference magnitude between the two groups was not big enough and hence did not reach statistical significance. Given the current pattern of results, we thus conclude that consolidation of the interfering knowledge does not increase interference strength and that the interference effects that we observed, both in naming latencies and accuracy (in Experiment 2), emerge immediately after or during learning, and do not appear to need time to evolve.

As with any null result, the question remains whether consolidation really does not affect interference strength, or whether we simply failed to detect the difference in the current study set-up. From research on novel word consolidation it appears that while lexical integration starts immediately during or after learning (e.g., Lindsay

[10] Even consolidated list 1 material in RI studies always refers to material learned within the context of the experiment, one (or at most a few days) before interference induction, but never, at least to our knowledge, to material learned years before, outside of the experimental context, as in the current study.

& Gaskell, 2013), it can take multiple days or even weeks to complete (Takashima et al., 2006). It is, thus, possible that a consolidation time window of just one night was too short to result in big enough differences in interference magnitude between consolidated and unconsolidated Spanish words. Conversely, it may be the case that the Spanish words were very easy and fast to integrate into the mental lexicon given that they refer to already existing concepts. Maybe vocabulary learning then resembles schema-consistent learning via fast (cortical) mapping (e.g., Coutanche & Thompson-Schill, 2014; van Kesteren et al., 2010). If this is the case, new foreign language vocabulary might not need the slow consolidation process that is otherwise required and thus would not benefit much from overnight consolidation.

Finally, we also have to consider the possibility that the interference effects we observe are not actually caused by lexical competition at final test (alone), but (rather) via inhibitory dynamics during the learning of the Spanish words. The English words, having been learned and consolidated long before participants took part in our study, might have been interfering with the acquisition of their Spanish translation equivalents and hence might have been suppressed to allow for efficient learning of the Spanish words (see Anderson, 2003, for a more in-depth discussion of inhibition as a mechanism driving forgetting and Chapters 2 and 3 for how this might apply to the language context). Quite possibly then, it is this lasting inhibition that made it harder for participants to recall these English words, compared to English words which did not need to be suppressed because no Spanish translations were learned. If our effects are indeed caused by inhibition *during* learning, consolidation *after* the learning phase would have little extra effect and hence explain why we failed to find a difference between the two groups in Experiment 2. In this scenario, consolidation could only have an added effect enhancing the inhibition that arose during learning through extra competition at final test, which in line with our earlier arguments, would presumably only be possible after the Spanish words had time to consolidate. Whatever the explanation, we encourage future research into the role of consolidation, if any, in foreign language attrition.

4.4.4 | Reconciling Findings from Experiment 1 with Experiment 2

Finally, it might seem puzzling at first sight that we found no interference effect in accuracy in Experiment 1, while we did in Experiment 2, especially in the absence of an interaction with consolidation time. However, a statistical comparison of the two experiments suggests that this apparent difference is not statistically reliable. The same holds for individual comparisons of the group from Experiment 1 with either of the two groups from Experiment 2 (though see below for a more detailed discussion).

For the first of those comparisons, contrasting the group from Experiment 1 with the no-consolidation group in Experiment 2, we had hypothesized, if anything, to observe a smaller interference effect in the latter group. This hypothesis was driven by recent reports within the memory literature that retrieving a memory can destabilize it and make it especially prone to subsequent interference. A destabilized memory supposedly needs time to reconsolidate, which is a process that is believed to take up to multiple hours (Stickgold & Walker, 2005). Asking participants to retrieve the English words in a pre-test immediately before the Spanish learning phase, as was the case in Experiment 1, might have thus artificially increased our chances of finding interference effects. It turned out that this concern was unwarranted though. In naming latencies, the two groups showed comparable interference effects. In naming accuracy, neither of the two groups showed an effect; and if anything, the interference effect was numerically larger in the no-consolidation group than in the group from Experiment 1.

This result is reassuring. Our RT interference effects are robust and were replicated even after removing any chance for the pre-test of the English words (and hence their possible destabilization) to facilitate those effects. But how come we did not observe even the slightest indication for a smaller interference effect in the no-consolidation group? In humans, studies on reconsolidation have focussed on procedural memory (e.g., Hardwicke et al., 2016; Walker et al., 2003) and fear memory (e.g., Kindt & Soeter, 2013; Schiller et al., 2010, see Agren, 2014, for a review). Whether reactivation of a *declarative* memory also makes it labile and in need of subsequent reconsolidation is less clear (compare Forcato et al., 2009; Hupbach et al., 2007; Strange et al., 2010; Wichert et al., 2011). In fact, the study most comparable in set-up and materials to ours did not find evidence for such a testing-induced destabilization either: Potts and Shanks (2012) taught their participants a list of English-Swahili translation pairs on day 1. On day 2, participants learned a second list of English-Finnish translations, the learning of which was either preceded by a test of the English-Swahili word pairs, or not. On day 3, everyone was tested again on English-Swahili translation pairs. The authors found evidence for retroactive interference (forgetting of English-Swahili word pairs due to the learning of English-Finnish associations) only in the group that learned Finnish without having the re-test of Swahili words before. The group who learned Finnish words after being re-tested on Swahili words instead benefitted from this reminder, and only performed slightly worse than a third (control) group who did not learn Finnish words at all (and thus had no interference whatsoever). In other words, contrary to what one would expect based on reconsolidation theory, Potts and Shanks' research suggests that retrieval of vocabulary does not make it more susceptible to interference from subsequent learning of translation equivalents, but instead appears to be strengthening it, making it, according to Potts and Shanks, 'immune' to such retroactive interference. The authors attribute their finding to the fact that their 'reminder test' required active retrieval of the English-Swahili word pairs, as opposed to just restudying them passively (as is the case in most other studies on reconsolidation), and speculate that under these circumstances testing is beneficial rather than detrimental (in line with the 'testing effect', see Antony et al., 2017).

Even though none of the groups in our studies differed from one another statistically speaking, the pattern of results is compatible with that reported by Potts and Shanks (2012). The English pre-test was an active retrieval test, and, numerically speaking, we report the least evidence for interference in accuracy rates in Experiment 1, where the English pre-test and the Spanish learning took place in immediate succession. Separating the English pre-test from the Spanish learning phase numerically increased the interference effect (i.e., in the no-consolidation group in Exp. 2). Consequently, having the English pre-test immediately before the Spanish learning phase did not make the English words more prone to interference; if anything, it appears to have made them less prone to subsequent learning (though this needs empirical validation).

Finally, we report the numerically strongest interference effects in the consolidation group from Experiment 2, for whom both the pre- and post-tests were separated from Spanish learning by a night of sleep. As already discussed, there was no reliable difference between this group and the no-consolidation group. The consolidation group did however differ numerically from the group in Experiment 1. Again, this difference is not statistically significant and hence should not be over-interpreted. However, we encourage future research to follow up on this trend. It appears that the combination of letting new words consolidate and removing the English reminder test from the Spanish learning interference phase presents the most favourable circumstances for interference effects to emerge.

4.4.5 | Conclusions

The current study provides the first empirical evidence of the detrimental effects that learning words from a new language can have on remembering words from previously learned foreign languages. Multilinguals are thus not wrong in their perception that adding a language to their repertoire will, sooner or later, hamper access to other foreign languages, even when those languages were learned a long time ago and to a high level of proficiency. Our results, furthermore, suggest that these effects emerge immediately, and do not need time (or overnight consolidation) to evolve.

CHAPTER 5

Individual Differences in Foreign Language Attrition: The Role of Language Use After a Study Abroad

This chapter is based on:

Mickan, A., McQueen, J. M., Brehm, L., & Lemhöfer, K. (2020). *Individual differences in foreign language attririon: The role of language use after a study abroad.* Manuscript submitted for publication.

I would like to thank Fabian Schneider, Dennis Joosen, Katharina Menn and Orhun Ulusahin for help with the transcriptions, Laurel Brehm for statistical advice and Wilbert van Ham and Arvind Datadien, Maurice Deutekom and Kees van der Ver for technical support with the online set-up of the study.

Abstract

While recent laboratory studies suggest that the use of competing languages is a major driving force in foreign language (FL) attrition, research on ‘real’ attriters has often failed to demonstrate such a relationship. We addressed this issue in a large-scale longitudinal study, following German university students throughout a study abroad in Spain and their first six months back in Germany. Monthly, percentage-based frequency of use measures enabled a more fine-grained description of language use than in previous studies. L3 Spanish forgetting rates were indeed predicted by the quantity and quality of Spanish use as well as by L1 German and L2 English fluency. Attrition rates were furthermore influenced by Spanish vocabulary knowledge at the end of the study abroad and amount of Spanish experience prior to the study abroad, but not by motivation to maintain Spanish or non-verbal long-term memory capacity. Overall, this study highlights the importance of language use for FL retention and sheds light on the complex interplay between language use and other determinants of attrition.

5.1 | Introduction

In the globalized world we live in, more and more people move abroad for extended periods of time: the EU-funded Erasmus program alone has sent out more than ten million people to other countries since its establishment in 1987 (European Commission, 2020). Along with the cultural enrichment and personal growth that comes with experience abroad, most people hope to improve their foreign language (FL) skills. Maintenance of abroad acquired language skills, however, often proves difficult upon return to one's home country. Returnees tend to start losing their FL skills soon after leaving the foreign country. How come we forget foreign languages and what determines how fast and how much we do so?

Recent experimental research has shown that FL attrition can come about through the use (or the learning) of other languages, especially other foreign languages (e.g., Bailey & Newman, 2018; Chapter 2; Chapter 3; Chapter 4; Isurin & McDonald, 2001; Levy et al., 2007). In Chapter 2, for example, we showed that naming pictures in L2 English or L1 Dutch (though to a lesser extent) hampers subsequent access to recently learned L3 Spanish translation equivalents. Merely by having participants retrieve translations in another language, Chapter 2 thus succeeded at inducing attrition in the lab and in doing so, established language interference as one possible driving force in FL forgetting. If these lab-based results generalize to real life, language usage patterns during the attrition period should be crucial determinants of the rate and extent of target foreign language loss in natural attriters. Paradoxically though, this prediction has not been borne out in previous research of foreign language attrition 'in the wild' (see Mehotcheva & Mytara, 2019). Have previous studies with real attriters failed to accurately assess and document the role of language use in foreign language attrition, or does language use simply play a much smaller role in real life than in tightly controlled lab situations? The present study seeks an answer to this question. Supported by regular frequency of use ratings, we asked most importantly (though not exclusively) whether there indeed is a relationship between language use and foreign language forgetting in the first six months after a study abroad, and whether there is an accessibility trade-off between the languages that an attriter speaks.

5.1.1 | Longitudinal vs. Cross-Sectional Designs

Previous research on foreign language attrition in real life can be divided into longitudinal and cross-sectional studies. The longitudinal approach is the most intuitive way of studying the phenomenon. Here, researchers follow a group of attriters over a period of time, assessing their FL skills at regular intervals. This approach, however, is very time-consuming for the researcher. Furthermore, participants tend to drop out along the way, resulting in often small, under-representative participant samples (e.g., N = 5 in Mehotcheva, 2010; N = 2 in Tomiyama, 2008; N = 4 in Yoshitomi, 1999; though see Murtagh, 2003, and Xu, 2010, for exceptions). To circumvent these issues with data collection, many studies on FL attrition have used cross-sectional designs instead (e.g., Abbasian & Khajavi, 2010; Bahrick, 1984a,b; Hansen & Chen, 2001), or a combination of cross-sectional and longitudinal designs (e.g., Grendel, 1993; Mehotcheva, 2010; Weltens, 1988). In cross-sectional studies, groups with differing attrition lengths are compared to each other as well as to a baseline group of non-attriters (i.e., a group of comparable learners) of the foreign language. Although cross-sectional studies have provided valuable insights, statistical comparisons in such studies rely on the groups being matched on a multitude of factors, such as age, socio-economic status, FL learning context, and the level of FL proficiency reached prior to attrition onset. While matching on the first three might be feasible, variation in FL proficiency is difficult to control. Attriters who have reached different levels of proficiency in the foreign language are likely to attrite at different rates (e.g., Bahrick, 1984a,b; Murtagh, 2003), making it important to account for initial proficiency in interpreting later attrition. Cross-sectional studies, however, have no way of accurately estimating prior FL proficiency on an individual basis. For a thorough investigation into individual differences in foreign language attrition, we thus instead need more large-scale longitudinal studies with a pre-attrition baseline measure next to the attrition measurement itself. A comparison between performance on the dependent measure (e.g., a vocabulary or grammar test) at baseline and at the attrition measurement (i.e., after a period of disuse) can then provide accurate and fine-grained participant-specific forgetting rates, which can serve as the basis for individual difference analyses. The present study aims at providing such a large-scale longitudinal dataset.

5.1.2 | Language Use as a Predictor of Foreign Language Attrition

From a theoretical point of view, researchers in the domain of (foreign) language attrition unanimously agree that language use should be one of the key factors in language attrition. For vocabulary maintenance at least, continued use is understood to be necessary to keep the activation thresholds of words in a given language low, and use of languages other than the target language (i.e., competing lexical entries) is thought to increase these activation thresholds and hence complicate subsequent retrieval (Activation Threshold Hypothesis, Köpke, 2002; Paradis, 1993, 2004). Both theory and lab-based experimental evidence thus point towards a clear role for language use and interference in attrition, and yet, as Mehotcheva and Köpke (2019) summarize, the majority of studies that have investigated the role of language use report no (consistent) relationship between language use (of the target and/or other languages) and foreign language retention.

Obviously, in real life, using a foreign language is inversely related to using other languages: code-switching aside, you only ever use one language at a time, and the more time you spend using a given language, the less time you have left for using other languages. Hence, what studies care about the most is usually only target language use (though use of other languages is often asked for as well, see for example Mehotcheva, 2010). In a large-scale cross-sectional study on the retention of school-/university-learned L2 Spanish, for example, Bahrick (1984a,b) found none of four Spanish language use measures (reading, writing, listening and speaking) to correlate with Spanish retention. Similarly, Mehotcheva (2010) failed to find clear evidence for a significant relationship between Spanish language use and Spanish vocabulary retention in German and Dutch learners of Spanish in the initial months after a study period abroad. Likewise, German/Dutch and English use did not predict Spanish attrition either. An exception to this pattern of results is a study by Alharthi and Al Fraidan (2016), who found that L2 English internet usage by L1 Arabic participants was a good predictor of L2 English proficiency after a 15-month attrition period. Yet, even in the latter study, other target language frequency of use measures, which should intuitively be just as important (such as watching TV in English, reading books, or attending English FL courses), did not predict L2 proficiency either.

What causes the failure to document a consistent, beneficial role of target language use, or conversely, a detrimental role of non-target language use on retention? Bahrick (1984a,b) reasoned that there was not enough variance in terms of Spanish use in his sample: all of his participants used very little to no Spanish during the attrition period. Others have argued that frequency of use measures often fail to accurately describe and capture language use because they focus too much on quantity (e.g., absolute hours of use) as compared to the quality of the input (see Schmid, 2007). Another possible problem might be how language use is quantified. Especially in cross-sectional studies, such as the ones by Bahrick (1984a,b) and Mehotcheva (2010), participants are asked to judge their frequency of use in retrospect. Estimating the overall amount of exposure to a language for a long period of time (more than 12 months for some of the participants in Mehotcheva's study and up to 50 years in Bahrick's sample) is difficult and prone to over- or underestimation. Moreover, judgments are often given on relatively subjective scales. Mehotcheva (2010), for example, had her participants judge frequency on a 5-point-scale from 'very rarely' to 'very frequently'. Taking these two aspects and the above discussed baseline problem for cross-sectional studies together, it is thus possible that frequency of use measures in previous studies were simply too imprecise to predict changes (or group differences) in proficiency. In the present study, we instead collected multiple, percentage-based frequency of use measures at regular intervals to circumvent this issue (see next section for details).

5.1.3 | The Current Study

To assess the role of language use in foreign language attrition, we followed a large group (N=97) of L1 German learners of L3 Spanish over the course of a year, spanning both their one-semester-long study abroad in Spain as well as their first half a year back in Germany. For feasibility reasons and to reach a large number of potential participants, we tested participants online. To assess changes in Spanish proficiency over time, we administered an online picture-naming vocabulary test in Spanish just before the study abroad (T1), at the end of the study abroad (T2) and roughly six months post return to Germany (T3). Since this study is about forgetting, the analyses reported below concern changes from T2 to T3. The T1 measurement is only relevant for two secondary analyses, as explained further below. We chose to study productive vocabulary for maximal comparability with the above-cited lab-based language attrition studies, but also because productive FL skills are known to attrite first (e.g., Bahrick, 1984a, b), thus increasing our chances of observing attrition in the six-month attrition window. In the remainder of this paper, terms such as Spanish proficiency and Spanish attrition always refer to Spanish *lexical* proficiency and Spanish *lexical* attrition, respectively, even when we do not explicitly specify this every time. Next to the Spanish proficiency test, we administered fluency tests in German (L1) and English (L2) and a language background questionnaire at each session, as well as a short frequency of use questionnaire once every month in between.

To get representative and accurate frequency of use estimates, we opted for a more continuous and less abstract frequency of use measure than previous studies. Instead of asking our participants once in retrospect, we asked them to estimate their *current* frequency of use *once every month* during the attrition period. This resulted in about six frequency of use measurements to average over rather than one single measurement. As in Mehotcheva (2010), we asked for frequency of use indications in the target foreign language Spanish, as well as in L1 German, L2 English and any other languages. Rather than making judgments on a scale, we asked our participants to estimate the percentage of time they currently spent speaking (and listening, reading and writing) each language. The total for each of these four domains had to add up to 100%, such that a given percentage reflected, for example, the percentage of time someone currently spoke Spanish out of the total amount of time they spent speaking. Because these percentages were given relative to a participant's personal total amount of language use (i.e., their 100%) rather than some subjectively perceived notion of 'rare' and 'frequent', we hoped that they would provide a more reliable estimate than indications on a scale such as the one used by Mehotcheva (2010) or estimates given in hours or minutes. We reasoned that this, together with the fact that we averaged over multiple measurements, would provide us with a more accurate approximation of frequency of use for each language during the attrition period and would therefore maximize the ability to observe a relationship between language use and attrition.

As already noted, amount of target FL use and the use of other languages are inversely related, and so finding a positive relationship between Spanish use and Spanish retention entails finding a negative relationship between Spanish retention and the amount of use of all other languages. An interesting question, however, is whether it matters *which* other languages an individual speaks. The laboratory results from Chapter 2 suggest that other foreign languages interfere more than a native language. Next to asking whether more frequent Spanish use during the attrition period helps Spanish retention, we thus also asked whether the ratio of L1 German over L2 English use during the remaining time makes a difference. Does someone who predominantly uses German when they do not speak Spanish suffer less from attrition than someone who predominantly uses English?

Next, in order to investigate whether there is a trade-off in accessibility between languages, we had to move away from the pure, mutually dependent frequency of use ratings. To do so, we administered fluency tests in L1 German and L2 English and asked whether changes in Spanish proficiency would go hand in hand with changes in fluency (i.e., accessibility) in these other two languages. If, as lab studies suggest, there is such a trade-off (caused through interference between languages) and assuming that frequent language use results in higher fluency scores, we should observe increases in fluency in German and English to be associated with proficiency decreases in Spanish.

Finally, since quantity of input in a certain language might not be the sole, or even most important predictor of Spanish retention, we also asked for the type of input our participants received. Next to asking for frequency indications in four domains (covering both productive language use, i.e., speaking and writing, and receptive language use, i.e., listening and reading), we also asked our participants to report what percent they received native as compared to non-native input in Spanish. Individuals who received solely native input, regardless of the total amount of input they get, might show less signs of attrition, or attrite more slowly than people who received mostly non-native and thus potentially incorrect input.

5.1.4 | Other Determinants of Foreign Language Forgetting

Language use is unlikely to be the only relevant predictor of forgetting rates 'in the wild'. Many factors have been implicated in foreign language attrition (for the most recent summary, see Mehotcheva & Mytara, 2019). In a naturalistic experiment, these other factors should be taken into account if one wants to arrive at a parsimonious model of foreign language attrition. In the current study, we thus additionally included a number of variables that either have been consistently shown to impact the rate of forgetting, or that have yielded contradictory results. In accounting for those variables, we hoped to get a more complete picture of the determinants of foreign language attrition.

5.1.4.1 | *Motivation*

Just as with language use, and in fact tightly linked to it, one might expect motivation to learn a foreign language well and one's attitude towards the FL to be important in determining how well FL skills are maintained. Once again though, the empirical evidence for a role of motivation in language maintenance (and conversely loss) is sparse (see Mehotcheva & Mytara, 2019, for a recent discussion). Just as with language use, previous studies assessed motivation only once, at the attrition measurement (i.e., once attrition had already occurred). Motivation before attrition onset, however, is arguably at least as important in determining the effort someone will put into maintaining a foreign language. Moreover, motivation is dynamic and can change over time (Nikitina & Furuoka, 2005), and hence can differ before and after attrition onset. One single motivation measure may thus not be an accurate reflection of someone's motivation throughout the complete attrition period. For a more nuanced and complete picture, we administered a shortened version of Gardner's Attitude and Motivation questionnaire (see section 5.2.2.4.2) at each of the main measurement sessions. We then averaged over the pre-attrition (T2) and attrition (T3) measurements to arrive at an estimate of overall motivation and attitude towards Spanish during the attrition period and asked whether this estimate predicted forgetting rates.

5.1.4.2 | *Amount of Experience with the Foreign Language*

Length of exposure to the foreign language and thus amount of experience with it prior to attrition onset is also often thought to be important. Usually, length of exposure is operationalized as the length of the stay abroad. In our case, all participants went abroad for only one semester. The variance in terms of study abroad length is thus minimal in our sample and most likely not meaningful in itself.

In a comparable population of German and Dutch learners of Spanish, Mehotcheva (2010) also found no conclusive evidence for a role of study abroad length on Spanish language retention, even though in her sample the range of the study abroad period was almost twice that in our sample. What is more, in adult FL learners who only go abroad for a short period of time, the study abroad is often only the tip of the iceberg. For many exchange students, much of the learning of the FL happens before they go abroad. It is hence variation in the amount of experience with Spanish prior to the study abroad that we think is the most important variable for the specific sample tested in our study. Our learners started their study abroad with different amounts of Spanish experience and hence we asked whether people with more years of experience were less prone to undergo attrition after the study abroad than those who had started learning Spanish more recently.

5.1.4.3 | *Foreign Language Proficiency Before Attrition Onset*

A variable that is closely linked to the amount of FL experience is the proficiency level reached in the FL prior to attrition onset. In fact, even though it has been claimed to be less important than amount of FL experience (Hansen, 1999), it is the variable that has been linked most consistently to forgetting rates. People with higher levels of FL attainment before attrition onset have repeatedly been shown to suffer relatively less from attrition than participants with lower levels of pre-attrition proficiency (e.g., Bahrick, 1984a,b; Mehotcheva, 2010; Murtagh, 2003). Bahrick (1984a,b), for example, found that participants who had followed more FL courses and who had received higher course grades prior to attrition onset maintained a higher proportion of vocabulary in what he called 'permastore' (i.e., vocabulary that remains available to a language user even after more than 25 years of disuse of the FL). While intuitively FL attainment should correlate with the amount of FL experience, this need not necessarily be the case. Some participants might be more efficient and faster learners than others and hence achieve higher levels of proficiency in a shorter amount of time. It thus seemed important to include both FL experience and FL proficiency in our analysis. We operationalized Spanish proficiency prior to attrition onset as performance on the Spanish vocabulary test at T2 (i.e., at the end of the study abroad and thus before potential attrition onset).

5.1.4.4 | *Long-Term Memory Capacity*

A factor that we know very little about in relation to (foreign) language attrition is general long-term memory capacity. Most recent theoretical accounts of (FL) attrition draw links between language and domain-general forgetting, and highlight the possibility of common underlying neural substrates and cognitive processes (see for example, Ecke, 2004; Köpke & Keijzer, 2019; Linck & Kroll, 2019; Mickan et al., 2019). If the same or at least similar mechanisms underlie both domain-general (i.e., non-verbal) and language forgetting, it may be that someone's rate of FL attrition is partially determined by their (non-verbal) long-term memory capacity. We tested this by administering a standardized visual long-term memory test at T3 (the Doors test, Baddeley et al., 1994, 2016). To the best of our knowledge, we are the first to investigate whether non-verbal long-term memory capacity predicts foreign language attrition severity.

5.1.4.5 | *Attrition Self-Judgment*

Finally, we ask whether participants have a realistic perception of how much they forget. Previous research suggests that participants tend to overestimate their personal amount of attrition (e.g., Murtagh, 2003; Weltens, 1988). While this is not a variable that is thought to predict or cause attrition, it is nevertheless interesting to ask whether we observe a similar overestimation of foreign language loss in our population.

5.1.4.6 | *Item-Specific Factors: Word Frequency and Cognate Status*

Next to individual difference factors, there are also variables that may influence the rate of forgetting on the item level. Cognates, for example, tend to be remembered better than non-cognates (e.g., de Groot & Keijzer, 2000; Weltens, 1988, though see Engstler, 2012). Likewise, there is some evidence that high frequency words are retained better than low frequency words (e.g., Mehotcheva, 2010; but see de Groot & Keijzer, 2000). We therefore included both of these factors in the analysis as well.

5.1.5 | Learning Context, Age and Other Constants in the Present Study

Of course, the above discussed variables do not constitute an exhaustive list of the factors contributing to foreign language forgetting. Including all possible predictors in one model was beyond the scope of the current study, and so, in an effort to reduce the number of predictors, we kept some factors constant. These factors include the learning context (natural / immersion or instructed), the languages involved, as well as the age and socio-economic status of the attriters and the length of the attrition period (see Mehotcheva & Köpke, 2019, and Mehotcheva & Mytara, 2019, for overviews and discussions of these variables). We selected our participants to be as closely matched on these aspects as possible: all of our participants were German university students between 20 and 29 years of age. They all went to Spain on a study abroad and hence learned Spanish both under natural circumstances while abroad as well as in a formal (classroom) setting prior to the study abroad. We scheduled the attrition session to take place roughly six months after the end of the study abroad for everyone, and participants were all back in Germany and immersed in their L1 at that time point. Six months is a relatively short attrition period. Previous research, however, suggests that most forgetting happens within the first months and years after attrition onset (Bahrick, 1984a,b). We were thus confident that we would observe some attrition within six months, and investigating a much longer period than that was simply not possible given the time constraints of the PhD project that this research was part of.

5.1.6 | Attrition as the Mirror Image of Acquisition?

Our design -and specifically the fact that we have a pre-study-abroad baseline (T1) next to the pre-attrition baseline (T2)- gave us the opportunity to ask an additional question about the nature of attrition, namely whether the forgetting process is the reverse of the acquisition process, and hence whether the words acquired last are also the first to be forgotten. This hypothesis was originally formulated by Jakobson (1941) to explain pathological L1 loss and is commonly known as the Regression Hypothesis (RH). Since its initial formulation, alternative versions have been advocated and researchers have argued that it might not be the information learned last, but rather the information learned least well that is forgotten first (e.g., Hedgcock, 1991). The jury on this debate is still out, partially because these two versions of the RH are hard to tease apart in real life, where information learned last often is also learned least well. In fact, there is evidence in support of the RH in FL attrition in both of its formulations (e.g., Hansen, 1999; Hedgcock, 1991; Kuhberg, 1992; Olshtain, 1989). Most of this evidence, however, comes from studies with children and from the domain of syntax and morphology. For vocabulary and late adult L2 learners, to the best of our knowledge, the RH has only been tested (and confirmed) once for *foreign* language attrition, namely in a cross-sectional study by Wang (2010). There is thus still a need for more research on whether the words learned last (or least well) are indeed the first to be forgotten. Having not only a pre- and post-attrition (i.e., T2 and T3) measure, but also a pre-study abroad baseline (T1), enabled us to ask, in an additional analysis, whether the words learned most recently (i.e., during the study abroad: unknown at T1 but known at T2), are more likely to be forgotten by T3 than the words that were already known before the study abroad (i.e., known at both T1 and T2). This does not distinguish the two versions of the RH from one another, but it does test the RH in its original formulation.

5.1.7 | Summary

In a nutshell, the present study aims at furthering our understanding of the factors underlying foreign language attrition. In a large-scale longitudinal study, we followed a group of German learners of Spanish throughout their study abroad in Spain and their first half a year back in Germany. Using each participant's Spanish vocabulary knowledge at the end of the study abroad as a pre-attrition baseline, we asked what best predicts participant-specific changes in vocabulary knowledge as measured six months post return to Germany. First and foremost, we were interested in the impact of language use, both in terms of quantity and quality, on attrition rates. We hypothesized that participants who still used Spanish frequently when back in Germany would decline only minimally in performance on the Spanish vocabulary test (or possibly even increase) while those who use very little Spanish would suffer more vocabulary loss. Furthermore, based on previous lab research, we predicted that a decline in Spanish proficiency would go hand in hand with fluency increases in German and English, and that participants who use more German than English would suffer less from attrition than those that use more English. Next to language use, we additionally took factors such as motivation and prior FL experience into account to arrive at the most parsimonious explanatory model of FL forgetting. Finally, apart from asking what determines individual forgetting rates, we also tested whether the study abroad had any long-term linguistic benefits at all (i.e., whether Spanish vocabulary knowledge returned to T1 levels at T3 or not), and whether we could find evidence for regression in foreign language attrition.

5.2 | Methods

5.2.1 | Participants

During the summer months of 2018, we asked the international offices of 33 German universities to email all their students that were about to embark on a study abroad in Spain with an invitation to our study. The email contained a link through which participants could indicate their interest in participating. In total, we received 481 sign-ups. Out of those sign-ups, we invited 194 German native speakers for the experiment. We only invited people who had grown up with German as their only mother tongue. Furthermore, we selected participants based on their study abroad length (4-7 months), their study abroad start date (no earlier than mid-August 2018, no later than October 2018) and based on whether they were planning on attending university courses in Spanish or English while abroad (see Appendix D.1 for details).

Due to participant drop-out and technical difficulties with the online recording of audio responses in some of the tasks, only 99 of these participants contributed data to both T2 and T3 (see Appendix D.1 for details on drop-out rates). Next to exclusions based on data availability, we excluded an additional two participants because their recordings indicated typing noises on over 20% of the trials of the Spanish naming task at either T2 or T3, suggesting that they were not paying full attention to the task and may have been consulting online dictionaries. The remaining 97 participants form the dataset for all analyses reported in this chapter.

Participants from the final set (71 females) were between 20 and 29 years of age (*M* = 22.30, *SD* = 1.80), had normal or corrected-to-normal vision and reported no history of neurological impairment or speech related disabilities. For all participants, English was their first foreign language. At the time of recruitment, prior to their study abroad, all of them knew some Spanish, though to varying degrees. Proficiency self-ratings at each of the three main measurement time points, both for English and Spanish, can be inspected in Table 5.1. Other languages that participants knew included most prominently French and Latin. As in previous chapters, Spanish is referred to as L3, even though it was in fact L4 or even L5 for some of our participants. We stick to L3 for simplicity.

22 of the 97 participants studied Spanish / Latin American studies at their German home universities; the remaining participants came from all kinds of study programs, including the natural sciences, law, business and medicine. Figure 5.1A illustrates where participants' home universities were situated. Study abroad destinations within Spain varied and can be inspected in Figure 5.1B. Participants went to Spain for on average 5.14 months (*SD* = 0.72, range = 3.53-7.96). At the time of the attrition measurement (T3), participants had been back in Germany for on average 6.17 months (*SD* = 0.45, range = 4.9-7.7; see section 5.2.2.2 for details on session timings).

Participants took part on a voluntary basis and were reimbursed via bank transfer with €20 per completed time point, thus making for a total of €60 if they completed all three sessions. The study was approved by the ethics committee of the Faculty of Social Sciences, Radboud University (ESCW2016-1403-391).

Table 5.1
Participant characteristics.

	M	*SD*	*range*
	Spanish		
Spanish AoA	15.82	3.24	10-26
Amount of experience with Spanish (in years, at T1)	3.78	2.32	0-11
Spanish proficiency self-ratings	T1		
Speaking	3.70	1.58	1-7
Writing	3.65	1.38	1-6
Listening	4.26	1.44	1-7
Reading	4.39	1.25	1-7
	T2		
Speaking	4.49	1.15	1-7
Writing	4.34	1.18	1-7
Listening	4.91	1.20	1-7
Reading	4.99	1.04	1-7
	T3		
Speaking	4.30	1.29	1-7
Writing	4.09	1.15	1-6
Listening	4.78	1.20	1-7
Reading	5.08	1.03	1-7
	English		
English AoA	8.86	1.93	3-14
English proficiency self-ratings	T1		
Speaking	5.29	0.91	3-7
Writing	4.93	1.03	3-7
Listening	5.68	0.84	4-7
Reading	5.63	0.98	2-7
	T2		
Speaking	5.26	0.90	3-7
Writing	4.98	1.10	2-7
Listening	5.72	0.80	3-7
Reading	5.73	0.88	3-7
	T3		
Speaking	5.23	1.08	1-7
Writing	5.04	1.22	1-7
Listening	5.77	0.92	2-7
Reading	5.73	0.97	4-7

Note. AoA = age of acquisition. Proficiency self-ratings were acquired on a 7-point Likert scale (1 = very poor, 7 = like a native speaker).

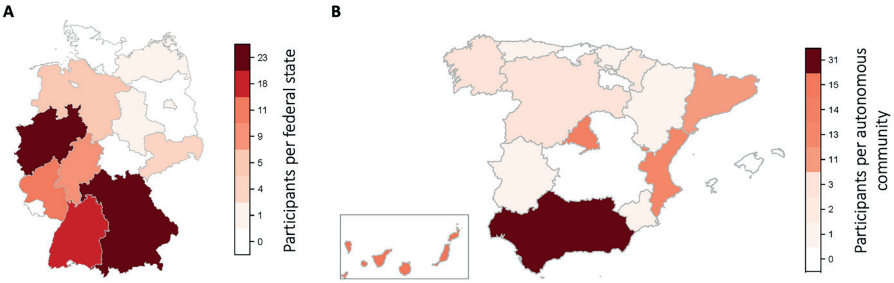
A
23
18
11
9
5
4
1
0
Participants per federal state
B
31
15
14
13
11
3
2
1
0
Participants per autonomous community

Figure 5.1

A Map of Germany showing where participants' home universities were located.
B Map of Spain showing where participants went to study abroad.

5.2.2 | Procedure & Materials

5.2.2.1 | *Overview*

We followed participants for one year, spanning their study abroad period as well as their first six months back in Germany, thus documenting both their study abroad and the initial stages of the subsequent attrition process. Throughout this time, participants were tested online at three time points: at the beginning of their study abroad (T1), at the end (T2), and roughly six months after leaving Spain (T3), when they were back in Germany (see Figure 5.2). We will subsequently refer to those measurement time points as 'sessions'. At each session, participants first completed a questionnaire, followed by a Spanish picture naming vocabulary test, and finally a number of English and German fluency tests. At T3 only, participants additionally completed a long-term memory test. In between these sessions, about once every month, participants were also asked to complete a questionnaire rating their current frequency of use in Spanish, English and German.

All tasks were administered online, but we stayed in touch with participants via email throughout the study. For each session, participants received an email with personalized links to each task, in the order they had to be completed in. They then had a maximum of two weeks to complete the tasks. If necessary, we sent reminder emails (see Appendix D.2 for details). For all tasks, participants were asked to find a quiet spot in which to do the tasks alone, within one sitting and in the indicated order. Time stamps on the logfiles and the audio quality suggest that all participants complied with those instructions. Logfiles and audio recordings were stored on a secure Radboud university server to which only the main experimenter had access.

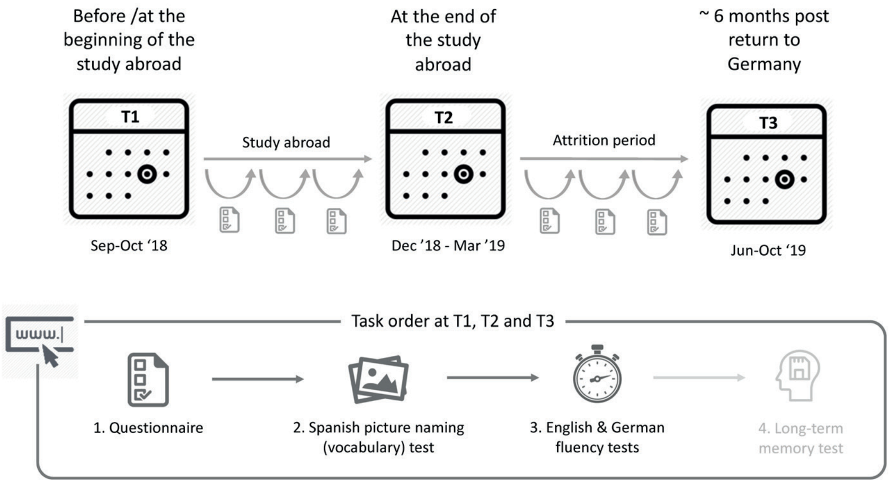
Before /at the beginning of the study abroad
At the end of the study abroad
~ 6 months post return to Germany
T1
Study abroad
T2
Attrition period
T3
Sep-Oct '18
Dec '18 - Mar '19
Jun-Oct '19
www.|
Task order at T1, T2 and T3
1. Questionnaire
2. Spanish picture naming (vocabulary) test
3. English & German fluency tests
4. Long-term memory test

Figure 5.2
Online study overview.

5.2.2.2 | *Timings of Sessions*

The timing of the tests was participant-specific and depended on when a participant had started to study abroad, when they were planning to return to Germany, and whether they had any extended trips to other countries planned in between, most notably trips back to Germany between T1 and T2 and trips back to Spain between T2 and T3.

5.2.2.2.1 | T1 Timing

As soon as participants signed up and were deemed suitable for the study (see section 5.2.1), they were invited for the first session. 10% of participants completed T1 before they left for Spain. The remaining participants completed T1 within their first weeks in Spain (45% within the first week, 38% within the second week, and 7% within their third week abroad).

5.2.2.2.2 | T2 Timing

T2 was initially scheduled to take place two weeks before participants left Spain. However, because the study abroad spanned the Christmas vacation and, as determined in a short pre-Christmas questionnaire, most participants went back to Germany for the holidays, we had to reschedule T2 for some of them. Participants who went back to Germany for a week or longer and who would return to Spain for only a relatively short amount of time (see Appendix D.3 for details), were invited to complete T2 before Christmas (29 out of the 97 participants). This was to ensure that each participant's T2 measurement reflected their peak Spanish performance as much as possible. Participants who still had a relatively long amount of time left in Spain after Christmas, or people who stayed in Spain for Christmas, were invited as originally planned, two weeks prior to the end of their study abroad. 77% of them completed T2 while still in Spain, 19% within the first two weeks back in Germany and 4% within the first three weeks back in Germany.

5.2.2.2.3 | T3 Timing

T3 was initially scheduled exactly six months after a participant left Spain. Given, however, that these T3 dates spanned the summer vacation period, we again sent around a questionnaire asking for participants' vacation plans. Based on participants' indications, we then adjusted individual T3 dates so that they would ideally take place before any vacation, most importantly before any trips to Spanish-speaking countries, while keeping the time between T2 and T3 maximal and close to six months. If this was not possible, we made sure that there was at least one month in between returning from vacation and the respective T3 test. On average, T3 took place 6.18 months after participants left Spain (*SD* = 0.45, range = 4.9 – 7.7).

Regardless of the timing, the tasks administered at each session always followed the same procedure. In what follows, we will describe each of the tasks in detail, starting with the dependent measure of the study, the Spanish proficiency test, followed by a description of the fluency tests and the questionnaires. Information from the latter two served as basis for the predictor variables used for modelling individual differences in Spanish proficiency.

5.2.2.3 | *Spanish Vocabulary Test (Dependent Variable)*

5.2.2.3.1 | Materials

At each session, participants named an identical set of 144 pictures of everyday objects and animals in Spanish (see the Appendix D.4 for a full list of items). The first four of those were practice items and were not included in analyses. Three additional items were excluded because their corresponding pictures turned out to be ambiguous (e.g., the picture of a 'pearl' often elicited 'ball'). Out of the remaining 137 experimental items, 19 were cognates between Spanish, German and English, 20 were cognates between Spanish and English only, and the remaining 98 words were non-cognates in the three languages. We defined cognates as translation equivalents with high form overlap. This included both identical cognates, such as 'sofa' (German: 'Sofa', Spanish: 'sofá'), and non-identical cognates, like 'botella' for the English word 'bottle' (see the Appendix D.4 for a full list). Because we did not see performance differences in Spanish recall between cognates in all three languages and cognates in only Spanish and English (neither at T2: $t(36.44) = 0.39$, $p = .699$, nor at T3: $t(36.91) = 0.54$, $p = .595$), we collapsed over the two types of cognates for the analyses below, distinguishing only non-cognates from cognates (of any type).

Experimental items were between one and five syllables long in Spanish ($M = 2.64$, $SD = 0.81$). Their Spanish log frequencies ranged from 1.08 to 4.67 ($M = 2.63$, $SD = 0.62$, according to the Spanish Subtlex, Cuetos et al., 2011) and their corresponding German log lemma frequencies ranged from 0.30 to 3.77 ($M = 2.29$, $SD = 0.65$, according to the German Subtlex, Brysbaert et al., 2011). We chose items from all frequency bands to make sure that participants' performance would not reach ceiling or floor at any session. One could argue for inclusion of either the Spanish or the German frequency counts in a model on Spanish forgetting rates. The two counts correlate highly ($r = .74$) and thus likely account for some of the same variance. Nevertheless, for analysis, we compared the two and chose the one which best predicted Spanish forgetting rates, which turned out to be Spanish log frequency (see Appendix D.5).

In order to make sure that our vocabulary test would be sensitive to the type of vocabulary knowledge acquired while abroad, item selection was furthermore informed by responses from a pilot study with five participants who had been on a similar stay in a Spanish-speaking country as our participants. For a bigger set of items, they indicated whether they knew each word and whether they thought they learned it during their stay abroad. We selected as many words as possible that these participants indicated to have learned abroad.

Pictures were photographs taken from Google images and the BOSS database (Brodeur et al., 2010). All pictures were displayed on a white background and occupied a maximum of 400 px in either width or length.

5.2.2.3.2 | **Procedure**

The task started with a microphone test. If the program could not detect a microphone, participants could not continue. If participants passed the microphone test, they were directed to an instructions screen. Participants were told that they would have to name pictures of objects and animals in Spanish and were asked to do so to the best of their knowledge, without consulting a dictionary and in a quiet room by themselves. They were instructed to say 'I don't know' if a Spanish label for a picture was unknown to them. Subsequently, the trial procedure was explained and four practice trials were administered. After that, participants could start the main part of the experiment via a button press.

Each trial started with the presentation of a picture in the center of the screen. The audio recording started automatically as soon as the picture appeared on screen. Participants then had a maximum of one minute to name the object or animal in the picture in Spanish. The recording stopped either after one minute or when participants clicked a button to indicate that they were done talking. The recording was then uploaded to the server, after which participants could proceed to the next trial. Each new trial had to be initiated by the participant via a button press. We chose this procedure to ensure that participants were still actively engaging in the task, and to enable them to refresh the website in between trials in case of internet connection issues. The order of presentation of the items was identical for all participants, but different between sessions T1, T2 and T3 (see the Appendix D.4 for session-specific item lists).

5.2.2.3.3 | Accuracy Scoring

As in Chapters 2 and 3, participants' Spanish productions were coded on the phoneme level. For each word we counted how many phonemes were produced correctly and how many were produced incorrectly. We chose a fine-grained coding because participants sometimes produced partially correct words (e.g., sella instead of sello; at T2: 5% in total and 14% of errors; at T3: 5% in total and 12% of errors), and using a dichotomous correct/incorrect coding would have ignored those nuances in the response variable (see also Chapters 2 and 3; de Vos et al., 2018). Incorrect productions could be either insertions, deletions or substitutions (see Levenshtein, 1966). Table 5.2 exemplifies the scoring procedure for the 'sella' example.

Table 5.2

Scoring example, phonetically transcribed.

Target word	s	e	ʎ	o
Participants production	s	e	ʎ	a
Scoring	correct	correct	correct	incorrect (substitution)

'Sella' would be counted as having 3 correct phonemes and 1 incorrect phoneme. Together these two numbers (3,1) formed the basis for the dependent variable for statistical modelling. For plotting and to provide descriptive statistics, we additionally calculated an accuracy percentage based on these two numbers. This percentage corresponds to the number of correct phonemes out of the total number of phonemes (e.g., for 'sella': (3/(3+1))*100 = 75%). When participants corrected themselves, or otherwise needed multiple attempts to name a picture, the last utterance was scored. Synonyms were counted as correct productions and their phoneme count was adjusted so that the total reflected the total phoneme count of the synonym.

5.2.2.4 | *Predictor variables*

5.2.2.4.1 | Fluency Tasks

Material. At each session, participants completed three German and three English fluency tests. Per language and session, we administered one letter fluency and two category fluency tests. The letters and categories differed across sessions and languages (see Table 5.3 and Appendix D.6 for details on selection procedure). All fluency tests were pre-tested with seven participants in German and English to ensure that they were at an appropriate level for speakers with low to moderate English proficiency.

Procedure. The fluency tests were administered in a fixed order that was identical for all participants: the English letter fluency task was followed by the two English category fluency tasks (in the order as displayed in Table 5.3), which were followed in turn by the German letter fluency test and finally the two German category tests. The fluency test also started with a microphone test, followed by instructions. Participants were told to name as many words as possible, but to avoid proper names, compounds with the same head (e.g., fish, fishnet), and inflections or derivations of words (e.g., run, ran, running, or actor, actors, actress). Each test started with a countdown of five seconds during which the participant saw the category or letter for the upcoming task accompanied by a British or German flag to make sure they knew which language to answer in. After this countdown, the recording started automatically and participants had exactly one minute to name as many words as possible belonging to the respective category or starting with the respective letter. After this minute, a pause screen appeared with a continue button that would start the next trial (i.e., the countdown) once participants clicked on it.

Table 5.3

Letters and categories chosen for the fluency tests.

	English		German	
	Letter	Categories	Letter	Categories
T1	F	land animals, professions	P	kitchen utensils, vegetables
T2	A	clothes, fruit	M	hygiene/bathroom supplies, transportation means
T3	D	body parts, electronic devices	B	sports, office supplies

Answer Scoring. Answers were coded offline. Each existing, unique word that fulfilled the restrictions described in the instructions above and that was a member of the tested category was counted as a valid word. The number of those valid words was used as the score for each task. Due to technical difficulties, the last fluency task in each session failed to completely upload to the server for some participants, such that recordings for 20% of participants took only 20-50 seconds rather than a full minute. Rather than excluding these participants, we therefore decided to omit the last fluency task. This means that for the German category fluency score, we only have one category (always the first listed per session in Table 5.3), rather than two. To arrive at a fluency score, we calculated difference scores for each language and separately for letter and category fluency. To do so, we subtracted the number of valid words for each given letter or category at T2 from their respective T3 scores and divided this number by the T2 score. The resulting difference scores thus indicate how much a participant's fluency increased or decreased with respect to their personal T2 baseline. For the English category fluency, we averaged over the two category scores; all other scores reflect performance on just one task per session.

Note that since the categories and letters differed per session, we do not know whether the tasks are comparable in terms of difficulty, and hence whether a numeric increase in fluency from T2 to T3 in our data truly reflects an increase in fluency after the study abroad. We thus have to leave the exact interpretation of mean fluency differences between T2 and T3 open. However, regardless of how comparable the tasks at T2 and T3 are, differences between participants are still meaningful, because all participants completed the same tasks and so should all equally be affected by differences in difficulty. Therefore, participants with the biggest relative fluency score increase from T2 to T3 (compared to other participants) will have maintained the best fluency, no matter whether T2 tasks are easier than T3 tasks or vice versa. The statistical models below thus ask whether a positive change in Spanish proficiency (relative to the other participants) goes hand in hand with (supposedly negative) changes in fluency in either English or German, and thus whether we can observe a trade-off in accessibility between languages.

5.2.2.4.2 | Questionnaires

At each session, participants filled in a questionnaire asking them for frequency of use ratings in all languages, proficiency self-ratings in Spanish and English, their motivation to learn Spanish, and a few session-specific questions about their study abroad and their time back in Germany. The full list of questions for each session in their original order can be inspected in Appendix D.7. Below we will only describe those parts of the questionnaires that we chose to include in the analyses reported below, sorted by relevance. Other questions were either not suited for analysis (open questions), not relevant for forgetting rates, or measured for outlier detection rather than analysis.

Frequency of Use Ratings. At each session, as well as once every month in between, we asked participants to estimate their current frequency of use in Spanish, German, English and other languages for the following four domains: reading, writing, speaking and listening. Estimates were given in percent and the sum of percentages for a given domain had to add up to 100% (e.g., Spanish 50%, German 20%, English 30%), such that a reported percentage reflected the relative time someone spent, for example, reading in a particular language out of the total amount of time they spent reading.

For analysis, we averaged over the four domains, resulting in one frequency of use percentage per language, session and participant. We chose to average over domains in an effort to reduce the number of predictors for analysis, and because correlations across domains at both T2 and T3 were high (all *r*'s > .6; see Appendix D.8 for a complete correlation matrix). We further averaged across all measurements *after* T2 (that is all in-between measures and T3), such that the final frequency of use percentage for each language reflected the average amount of time a participant spent using that language after their study abroad.

Given that the frequency of use indications for all languages together add up to 100%, they are mutually interdependent, and hence cannot be entered into a statistical model together. Because of this, we reduced the three frequency of use measures (for English, German, and Spanish) to two: Spanish frequency of use and the ratio of English to German use. The ratio reflects whether someone predominantly used German (L1) or English (L2) during the time not spent speaking Spanish. We included the ratio measure to answer whether or not it matters which other language you speak (when not speaking Spanish). Because we divided English frequency by German frequency, a ratio greater than 1 reflects more use of English compared to German, a ratio of 1 reflects equal use of the two languages, and a ratio between 0 and 1 reflects more use of German compared to English.

Motivation Questionnaire. At all sessions, we asked participants to fill in a motivation questionnaire. The questions were identical in all sessions, and were taken in part from Mehotcheva (2010) and in part from Gardners' Attitude and Motivation Test Battery (AMTB, Gardner, (1985).[11] All questions were translated into German, and were adjusted to the study abroad context when necessary (see Appendix D.7 for

[11] The questionnaire used by Mehotcheva (2010) was also based on the AMTB, yet was missing questions from one of Gardner's subscales (anxiety).

the full list of questions). The final questionnaire consisted of 37 questions in total with answers given on a scale from 1 ('completely disagree') to 7 ('completely agree'). Questions were divided over five subcategories (in line with Gardner's taxonomy) asking participants about their:

1 General motivation to learn foreign languages (N = 11)
2 Attitude towards the Spanish people (N = 9)
3 Integrative motivation to learn Spanish (i.e., for social & intrinsic reasons) (N = 6)
4 Instrumental motivation to learn Spanish (i.e., for pragmatic/utilitarian reasons, such as finding a job) (N = 5)
5 Anxiety / nervousness related to using Spanish (N = 6)

For analysis, we calculated averages for each subcategory and participant. Before averaging, answers to negatively formulated questions (25% of questions) were reversed. For modelling, we followed Gardner's suggestion to average each participants' scores on the first three subcategories to arrive at an overall score of integrative motivation per session. This choice was reinforced by the fact that scores on these three categories correlated highly positively with one another (all *r*'s and T2 and T3 > .5, see Appendix D.9 for the full correlation matrix), while scores on instrumental motivation and anxiety did not correlate strongly with any of the other categories (*r*'s < .3). We thus kept the latter two subscores separate.

Finally, because we were interested in how someone's *average* motivation to learn Spanish after the study abroad would affect their Spanish vocabulary development from T2 to T3, we chose to average T2 and T3 scores for each of the three subcategories for each participant. Someone who loses motivation from T2 to T3 would thus have a somewhat lower overall motivation score than someone who stays highly motivated. Note that the T2 measurement for most participants took place while they were still in Spain, and is hence not strictly speaking part of the time interval we are interested in here, but because we do not have monthly measures for motivation (as we do for frequency of use), including T2 for this measure comes closest to estimating overall average motivation for the attrition period (which using only T3 would not). Note also that another option would have been to calculate difference scores and to quantify the change in motivation from T2 to T3. This seemed less appropriate though given that such a measure would equate highly motivated participants that stayed highly motivated with poorly motivated participants that stayed poorly motivated.

Type of Spanish Input. At both T2 and T3, we asked participants whether they were currently regularly speaking Spanish, and if so, to what percent they were doing so with Spanish native speakers as compared to non-native Spanish speakers. The total percentage had to add up to 100%. This indicates what percentage of their Spanish input came from native speakers, *regardless of the total amount of Spanish input they received*. Since this question depended on participants answering that they were still actively using Spanish, at T3 we only have answers to the former question from 47 out of the 97 participants. Hence, we could not include this variable in the main analysis. We did, however, run an additional analysis including this variable on this subset of 47 participants. For this secondary analysis, we used the average percentage of native input across T2 and T3 as predictor for Spanish forgetting rates.

Attrition Self-Judgment. Finally, at T3, we asked participants to judge whether their Spanish had improved or worsened since returning from Spain. Judgments were made on a scale from 1 (worsened a lot) to 7 (improved a lot), with 4 reflecting no (perceived) change. The resulting score reflects how much participants thought they attrited from T2 to T3 and was entered into the statistical models as is, in order to find out whether participants' own judgments align with our objective Spanish proficiency measure.

5.2.2.4.3 | Doors Test

As a last task, and at T3 only, we asked participants to complete the 'Doors test', a visual long-term memory test developed by Baddeley et al. (1994). Out of the 100 target-foil sets available on the 'Doors of memory' website (https://www.york.ac.uk/res/doors/resources.shtml; Baddeley et al., 2016), we chose 30 sets to make the test short enough to administer online (see Appendix D.10 for a list of doors). The test itself started with an encoding phase, in which participants saw each of 30 target door pictures once for one second (we followed the procedure detailed in Baddeley et al., 2016). Participants were told to remember the doors and that they would later be tested on them. Between pictures, participants saw a blank screen for 200 ms. Pictures were presented automatically, without requiring a response from the participant. This encoding phase was followed by the test phase, in which participants saw 30 picture assemblies, consisting each of one previously seen door (i.e., the target) and three foils, matched in style and color to the target door (see Baddeley et al., 2016, for details on foil selection). The participants' task was to select the target door by clicking on it. There was no time limit and participants did not get feedback on their selection. For analysis, we calculated the percentage of correctly recognized target doors in the test phase.

5.2.2.4.4 | Overview of Predictors

Based on the questionnaires and tasks described above, we chose to assess 16 variables as predictors of forgetting rates from T2 to T3. Table 5.4 summarizes all variables and how we arrived at them. Next to the predictors described in detail above, we also included amount of experience with Spanish prior to the study abroad and T2 performance as predictors (see Introduction, section 5.1.4, for motivation), as well as the item-level predictors word frequency and cognate status.

Table 5.4
Overview of predictor variables assessed via model comparison (German = L1, English = L2, Spanish = L3).

Participant-level predictors	
Predictor name	**Description**
Frequency of use	
Spanish	average frequency of use of Spanish (in %), averaged over the four domains (speaking, writing, listening, reading) for the entire period after returning to Germany (i.e., all measures taken after T2)
German-English ratio	ratio of German over English use, averaged over the four domains for the entire period after returning to Germany. score = 1 -> equal use of German and English score < 1 -> more German than English score > 1 -> more English than German
Fluency	
English letter	relative difference in fluency between T2 and T3 ((T3-T2)/T2) on the English letter test
English category	relative difference in fluency between T2 and T3 ((T3-T2)/T2) on the average over both English category tests
German letter	relative difference in fluency between T2 and T3 ((T3-T2)/T2) on the German letter test
German category	relative difference in fluency between T2 and T3 ((T3-T2)/T2) on the first German category test
Motivation[12]	
Integrative	average integrative motivation score at T2 and T3
Instrumental	average instrumental motivation score at T2 an T3
Anxiety	average anxiety score at T2 and T3
Memory capacity	% correct score on the Doors test
T2 Spanish proficiency	% correct in the Spanish vocabulary test at T2
Amount of experience with Spanish[13]	amount of experience with Spanish prior to the study abroad (in years), as reported in Table 5.1
Attrition self-judgment	subjective rating of having improved or gotten worse in Spanish since returning from Spain, indicated on a 7-point Likert scale
Amount of native Spanish input	average % of native Spanish input (regardless of the total amount of Spanish input) across T2 and T3
Item-level predictors	
Word frequency	Spanish log frequency
Cognate status	distinguishing non-cognates from cognates (in any of the three languages, e.g., German-Spanish-English or English-Spanish)

[12] To reduce the number of predictors, we only included the instrumental motivation score in the final model. We chose this score (over the other two motivation scores) based on model comparison (i.e., the lowest AIC and BIC; see Appendix D.5).

[13] Amount of experience with Spanish was a significantly better predictor of forgetting rates than study abroad length ($p < .001$).

5.2.3 | Modelling

To investigate individual differences in forgetting rates, we ran logistic mixed effects models in R (Version 3.5.1, R Core Team, 2018), using the lme4 package (version 1.1-21, Bates et al., 2015) and the optimizer 'bobyqa'. As in Chapters 2 and 3, the dependent measure for all these models was the odds of correctly producing a phoneme for a given target word in the Spanish vocabulary test. A two-column matrix with the number of correct and incorrect phonemes for each target word at both T2 and T3 was passed to the model as dependent variable (this is one of multiple ways of specifying the response variable in binomial models, see https://www.rdocumentation.org/packages/stats/versions/3.2.1/topics/family); see also Chapters 2 and 3 and de Vos et al., 2018).

Forgetting rates were indexed by a change in accuracy from T2 to T3. We were interested in declines (or improvements) in Spanish after returning from abroad, and what predicts these changes in proficiency. To that end, we included a variable that coded for whether a data point belonged to T2 or T3, the so-called Session variable. The Session variable was effects coded (-0.5, 0.5), such that a negative beta estimate for this variable reflects forgetting (i.e., a decrease in accuracy from T2 to T3) and a positive estimate reflects learning (i.e., an increase in accuracy from T2 to T3).

In order to ask whether any of our above-listed predictors modulate forgetting (or learning), we entered each predictor in interaction with Session into the model. A significant interaction term means that a predictor has a modulating effect on the change in accuracy from T2 to T3. A positive estimate reflects that with every unit increase in the predictor variable, the difference between T2 and T3 increases, with T3 accuracy exceeding T2 accuracy; which translates to a learning effect that increases with an increase in the predictor variable. A negative interaction estimate, in turn, reflects that with every unit increase in the predictor variable the difference between T2 and T3 also increases, however, in the opposite direction, with T2 accuracy exceeding T3 accuracy, which translates to a forgetting effect that increases with an increase in the predictor variable.

For each of the above listed predictors, we checked whether their inclusion in a model with Session significantly improved model fit compared to a baseline model with Session as the only predictor. If inclusion of a predictor indeed improved model fit, as assessed via chi-square model comparisons, the predictor was later included in the final full model, otherwise it was discarded (see Appendix D.5 for model comparison outcomes). In the final model, we then entered all significant predictors, each in interaction with the Session variable, together. We did this in order to reduce the number of predictors in the final model and to arrive at the most parsimonious final model justified by the data. All models, both the separate initial models as well as the final full model, included random intercepts for both Subject and Item, and all p-values were calculated by model comparison, using chi-square tests, omitting one factor at a time.

Next to the main analysis, we also had three secondary research questions (see Introduction, section 5.1). To make the results section easier to follow and to avoid repetition, we will describe the details of the statistical models run to answer these additional questions only in the respective subsections of the results section.

5.3 | Results

5.3.1 | Descriptive Statistics for the Dependent Variable and the Predictor Variables

5.3.1.1 | *Spanish Picture Naming Performance*

Figure 5.3 shows participants' performance in the Spanish vocabulary test at T1, T2 and T3. On average, participants learned Spanish words while abroad (absolute learning rate between T1 and T2: *M* = 17%, *SD* = 8%, range = -1% – 42%) and forgot words after returning to Germany (absolute forgetting rate between T2 and T3: *M* = 4%, *SD* = 6%, range = -12% – 21%). However, as the by-participant data show (plotted in light- and dark-grey lines in the background), there is a lot of variation. Zooming in on T2 and T3, it turns out that while the majority of participants forgot some Spanish, some forgot much more than others, and some participants even improved from T2 to T3. It is these individual differences that we hope to explain with the above-listed predictor variables.

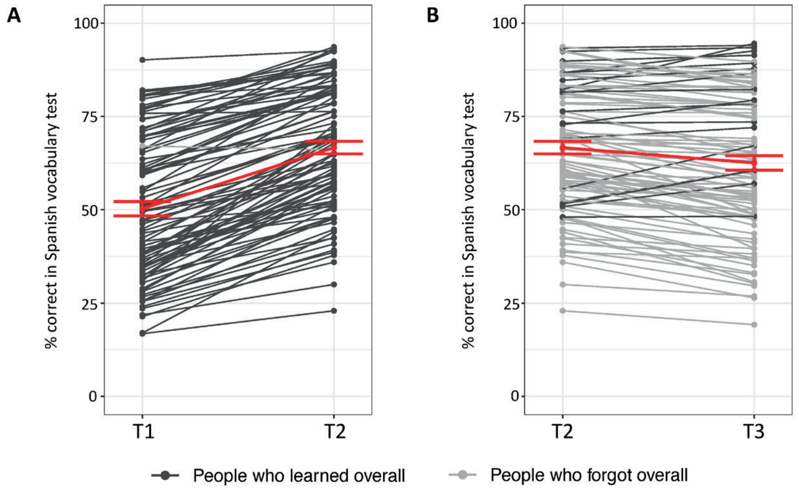

Figure 5.3
A. Participants' performance on the Spanish vocabulary test at T1 and T2. Dark grey lines represent people who learned on average while abroad, light grey lines represent people who forgot while abroad (N = 1). **B**. Participants' performance on the Spanish vocabulary test at T2 and T3. Dark grey lines reflect participants who learned on average after returning from abroad, light grey lines reflect participants who forgot after returning to Germany. In both subplots, the red line reflects the group mean and error bars correspond to the standard error around the session means.

5.3.1.2 | *Frequency of Use*

Figure 5.4 shows how frequency of use changed throughout the duration of the experiment (panel A), as well as how the resulting two frequency of use predictors (for Spanish, and the ratio between English and German use) for the time after T2 are distributed (panel B). After leaving Spain, participants spoke German the vast majority of the time and very little Spanish and English. The rather narrow standard error around the mean in panel A, as well as the corresponding histogram in panel B, furthermore shows that there is relatively little variability in the frequency of use measures, especially with regard to the ratio of English to German use. None of our participants used more English than German (values are never above 1) and the majority used far more German than English. For Spanish use, the distribution is also right-skewed and most participants used Spanish 10% of the time or less.

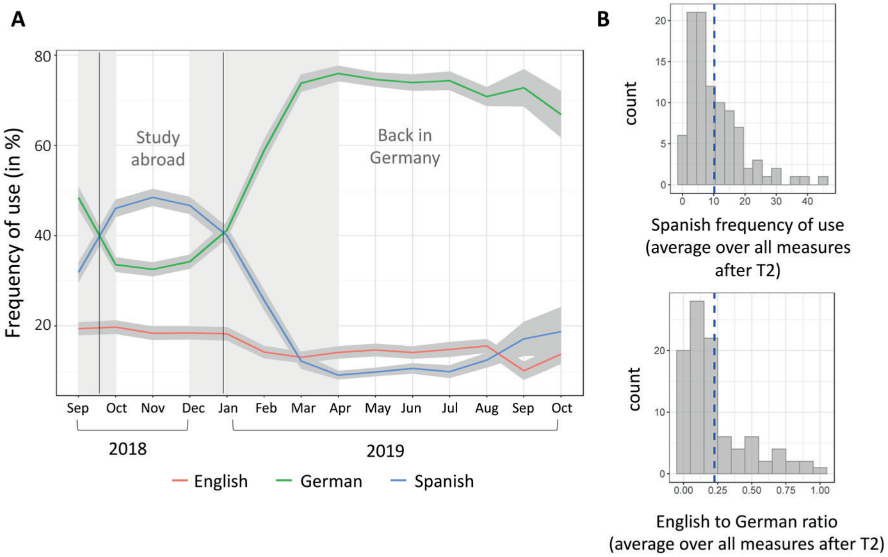

Figure 5.4

A. Average frequency of use for each language throughout the duration of the study abroad as well as their time back in Germany. Grey areas reflect the standard error around the mean. Vertical stripes indicate the average start and end date of the study abroad and grey areas around those averages reflect the absolute ranges of start and end dates respectively. **B**. Histograms for the two frequency of use predictors for modelling. The dashed blue line reflects the mean for each variable.

5.3.1.3 | *Fluency*

Participants' performance on the fluency tasks at T2 and T3, as well as histograms for the resulting predictors can be inspected in Figure 5.5. Because the tasks were different at each session, absolute changes from T2 to T3 are not directly interpretable. We are instead interested in whether relatively large or small changes in (German and English) fluency scores from T2 to T3, compared with the other participants, predicted Spanish forgetting rates.

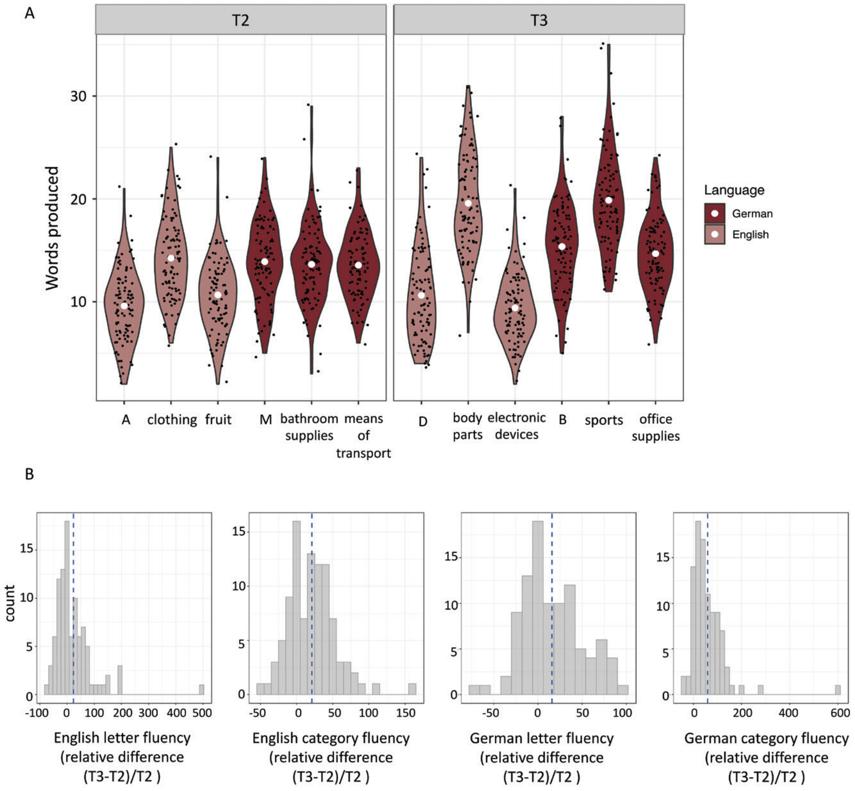

Figure 5.5

A. Violin plots for the number of words produced in each of the fluency tasks at T2 and T3. Categories/letters are plotted in the order that they were administered in. White dots represent means per task. Black dots represent individual participants. Violin plot outlines represent the distribution of the data. **B**. Histograms for the four resulting predictor variables used for modelling. Dashed blue lines reflect the mean for each predictor.

5.3.1.4 | *Motivation*

Average motivation scores for each subcategory at T2 and T3, and the distributions of the corresponding predictor variables, can be inspected in Figure 5.6. Participants were, overall, very motivated to learn Spanish, and were so more out of personal interest and affinity with the language than for practical reasons (compare integrative with instrumental motivation). There was also much less variability in participants' integrative motivation compared to their instrumental motivation and their anxiety to speak Spanish, as can be seen from the individual scores plotted in grey in the background. Moreover, on average, participants' motivation, both instrumental and integrative, as well as their anxiety to speak Spanish changed minimally from T2 to T3. Again, there was considerable variability between participants though. As explained above, for analysis, we averaged across T2 and T3 for each participant and each motivation subscore. In doing so, we are approximating each participant's average motivation throughout the time period under investigation. As explained in Table 5.4, only the instrumental motivation score entered the full model reported below. Including all three motivation scores was not possible with the current sample size, and we chose instrumental motivation because (as confirmed via model comparison) it was the best predictor of forgetting rates out of the three subscores.

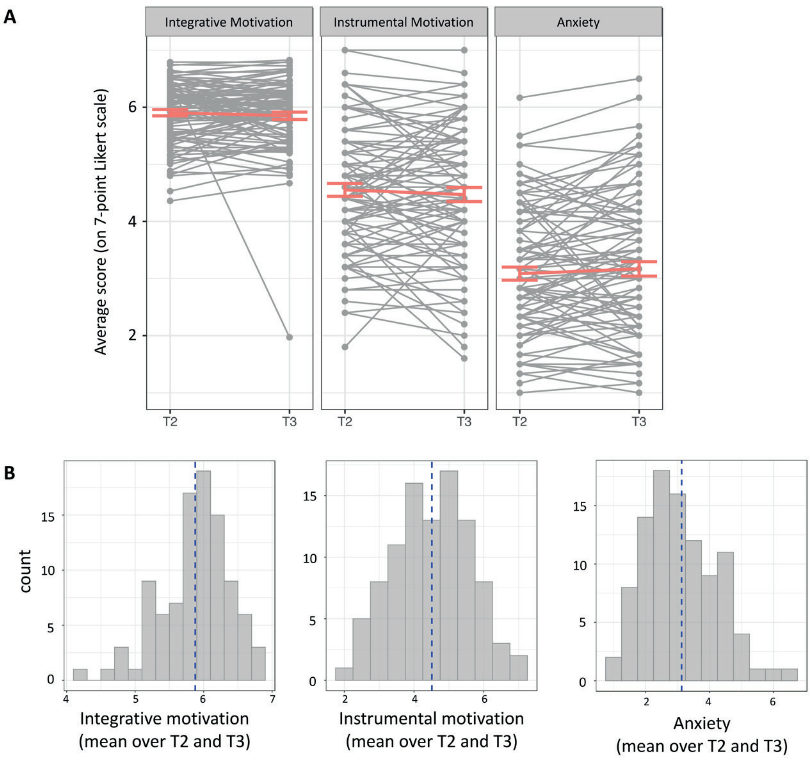

Figure 5.6

A. Average scores on each of the three subparts of the motivation questionnaire at T2 and T3. Light grey lines and dots reflect participant averages. Red lines reflect the means with the error bars denoting the standard error around the mean. **B**. Histograms for the three resulting predictor variables. The dashed blue lines reflect the mean for each predictor.

5.3.1.5 | *Remaining Participant-Level Predictors*

Distributions for the remaining predictors can be inspected in Figure 5.7. Panel A shows the memory performance measure. Performance on the Doors test was low but on average above the chance level of 25%. With the exception of a few outliers, participants recognized on average only half of the 30 doors from the encoding phase (*M* = 53%, *SD* = 15%). Panel B shows the average performance in the vocabulary test at T2. T2 performance varied, but not a single participant was at floor or at ceiling. On average, participants knew 66% of the words on the vocabulary test (*SD* = 17%, range = 23%- 94%). Panel C shows self-judgments of attrition. The average participant thought their Spanish had not changed after moving back to Germany (*M* = 3.92, *SD* = 1.57). Again though, there is considerable variation in participant's attrition self-judgments with answers spanning almost the full scale of answers (from 1 'got a lot worse' to 7 'got a lot better').

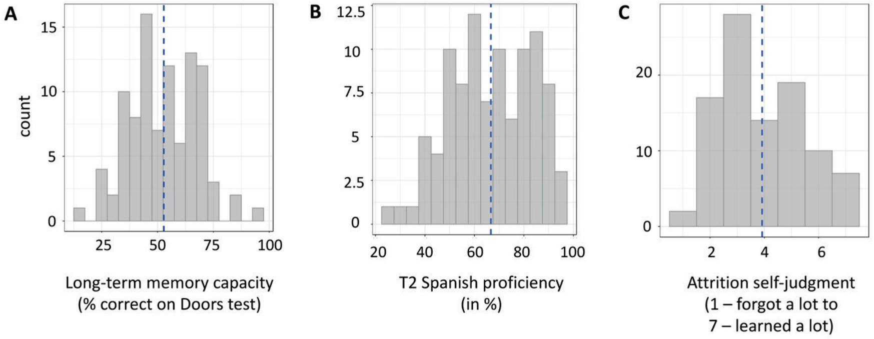

Figure 5.7
Histograms for the remaining predictor variables used for modelling. The dashed, blue lines reflect the mean for each predictor.

5.3.1.6 | *Between-Predictor Correlations*

A correlation plot for all 13 participant-level predictors is shown in Figure 5.8. The strongest associations exist between Spanish frequency of use and T2 performance (*r* = .51) and Spanish frequency of use and the motivation questionnaire scores: integrative (*r* = .36) and instrumental motivation to learn Spanish (*r* = .49) and anxiety to speak Spanish (*r* = -.44). T2 performance also correlated positively with integrative motivation (*r* = .42) and negatively with anxiety to use Spanish (*r* = -.44). No correlations exceeded .51 and all variance inflation coefficients are below 1.8, indicating that our predictors are sufficiently independent to be included within one explanatory model. Out of all 15 predictor variables (including the two item-level predictors), ten made it into the final model (see Appendix D.5 for the results of the model comparisons and Table 5.5 below for the final full model outcome).

For the smaller subset of participants (N = 49) that we used for the analysis including the 'amount of native Spanish input' predictor, the same relationships between predictors hold (see Appendix D.11 for the corresponding correlation matrix). The Spanish input predictor does not correlate highly with any of the other predictors (all *r*'s < .22).

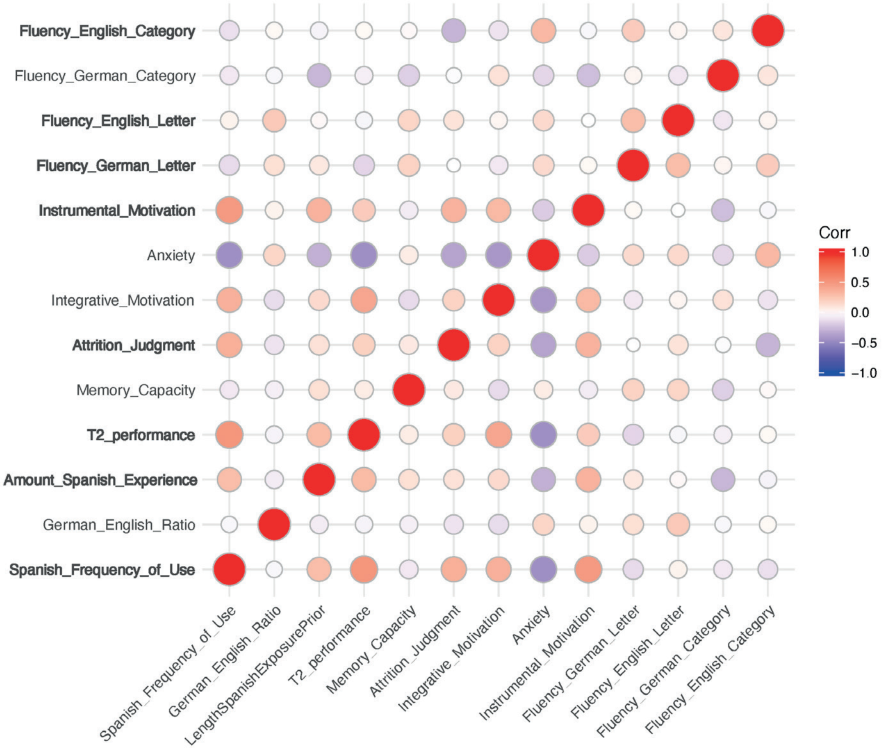

Figure 5.8

Pearson correlation matrix for all 13 participant-level predictors from the main analysis. Colors indicate the strength of the correlation (Pearson's r) with shades of blue indicating negative and shades of red indicating positive correlations. Predictors in bold made it into the final model.

5.3.2 | Regression Model Outcomes

5.3.2.1 | *What Predicts Forgetting Rates After a Study Abroad?*

In the main analysis we asked whether participants indeed forgot Spanish after returning to their home countries, and whether the extent to which they forget could be predicted by any of the participant- and/or item-level predictors discussed above. Model outcomes from the mixed effects logistic regression that we ran to answer these questions can be inspected in Table 5.5.

We will first discuss the participant-level predictors, and after that the item-level predictors. In both cases, we will only discuss significant main effects if the predictor did not also significantly interact with session. For predictors that significantly modulated forgetting rates, their relationship to Spanish performance is illustrated in Figure 5.9. First, we observed a main effect of Session such that participants made more errors in the Spanish vocabulary test at T3 than at T2. This means that after roughly half a year back in their home country, participants had overall indeed forgotten some of the Spanish they used to know at the end of their study abroad (*M* = 4%, *SD* = 6%). This main forgetting effect was modulated by a number of factors, including most importantly, Spanish frequency of use: people who still used Spanish relatively frequently after leaving Spain forgot less on average (and in fact even continued learning) compared to individuals who used less Spanish (see Figure 5.9 for visualization). English and German letter fluency also significantly modulated forgetting rates, and did so in opposite ways (Figure 5.9). People whose English letter fluency scores (relative to other participants) increased from T2 to T3 forgot less Spanish than people whose English letter fluency decreased.[14] For German letter fluency, we observed the opposite pattern: participants whose German letter fluency scores increased from T2 to T3 forgot more Spanish than people whose German letter fluency scores decreased.

[14] A closer look at the predictor plot shows that there is an extreme outlier in the English letter fluency task, which might be contributing disproportionately to the effect. Running the model again without this outlier, however, shows identical results.

Table 5.5

Logistic mixed effects model output for main analysis.

Fixed effects	Estimate	*SE*	*z*	$p(\chi^2)$
Intercept	**2.00**	**0.17**	**11.98**	**< .001**
Session	**-0.32**	**0.02**	**-15.21**	**< .001**
Participant-level predictors				
Spanish frequency of use	**0.10**	**0.03**	**3.07**	**.002**
Amount of experience with Spanish	0.04	0.03	1.42	.156
Attrition self-judgment	0.05	0.03	1.64	.100
English category fluency	-0.02	0.03	-0.73	.467
English letter fluency	0.01	0.03	0.49	.623
German letter fluency	-0.02	0.03	-0.78	.433
Instrumental motivation	-0.01	0.03	0.12	.843
T2 performance	**1.34**	**0.03**	**46.93**	**< .001**
Session * Spanish frequency of use	**0.21**	**0.02**	**10.07**	**< .001**
Session * Amount of Spanish experience	**0.07**	**0.02**	**3.89**	**< .001**
Session * Attrition self-judgment	**0.08**	**0.02**	**4.50**	**< .001**
Session * English category fluency	0.00	0.02	0.27	.788
Session * English letter fluency	**0.07**	**0.02**	**4.74**	**< .001**
Session * German letter fluency	**-0.05**	**0.02**	**-2.92**	**.004**
Session * Instrumental motivation	-0.01	0.02	-0.46	.643
Session * T2 performance	**-0.05**	**0.02**	**-2.66**	**.008**
Item-level predictors				
Word frequency	**1.02**	**0.15**	**6.70**	**< .001**
Cognate	**2.30**	**0.33**	**6.91**	**< .001**
Session * Word frequency	-0.03	0.02	-1.88	.060
Session * Cognate	**0.12**	**0.04**	**2.83**	**.005**
Random effects	Groups	Var	*SD*	
Item	Intercept	3.54	1.88	
Subject	Intercept	0.05	0.22	

Note. Critical parts of the table (i.e., main effect of session and interaction terms) are highlighted. Significant effects are marked in bold. *SE* = standard error; $p(\chi^2)$ = Chi-square p-value; Var = variance; *SD* = standard deviation.

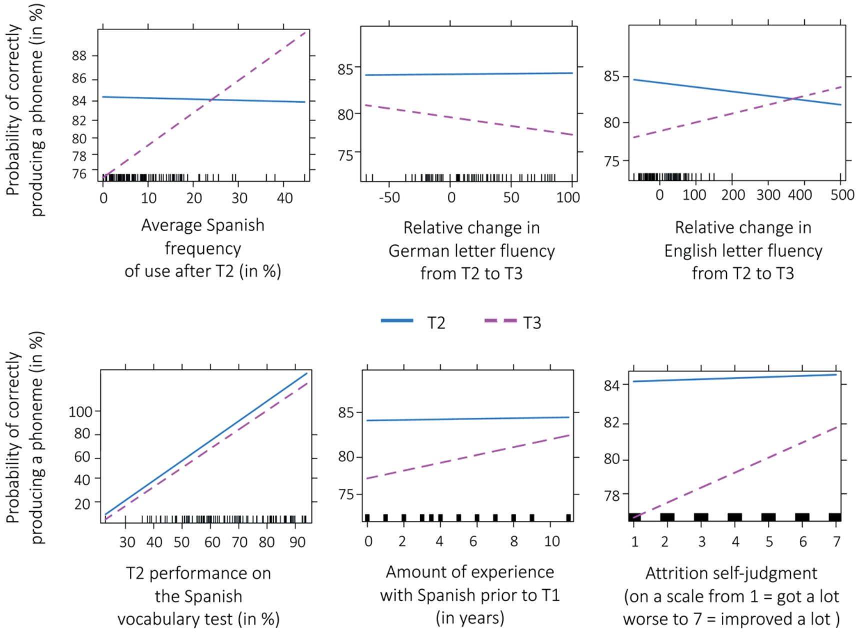

Figure 5.9

Participant-level predictor plots derived from the model outcome in Table 5.5 for all significant Session*Predictor interaction terms. The y-axis shows the probability of correctly producing a phoneme in the Spanish vocabulary test and is plotted on the logit scale (the link scale used for modelling) and labeled on the probability scale. The x-axes plot the respective participant-level predictors. Tick marks on x-axes reflect values for individual participants. Effects are plotted separately for T2 and T3. When the dashed, pink line is above the solid, blue line, a learning effect is predicted (better performance at T3 than at T2); conversely, when the solid, blue line is above the dashed, pink line, the model predicts forgetting (worse performance at T3 than at T2). Interactions concern changes in the difference between these two lines over different values of the respective predictor. For Spanish frequency of use (top left panel), for example, the model predicts that participants who use Spanish ~25% or more of their time learn from T2 to T3 while participants who use Spanish less than that tend to forget Spanish from T2 to T3. For T2 performance (bottom left panel), the opposite holds. Note that for T2 performance, the blue (T2) line runs perfectly diagonal, because average T2 performance by definition perfectly predicts T2 scores. Again, for the interaction though, the difference between the lines matters: the dashed, pink line increasingly diverges from the blue line, meaning that participants who performed better at T2 tended to forget more from T2 to T3.

Next to frequency of use and fluency, amount of Spanish experience prior to the study abroad also significantly predicted forgetting rates, such that participants with more Spanish experience before the study abroad forgot less than participants with less prior Spanish experience. Conversely, participants who performed better on the vocabulary test at T2, and who thus supposedly reached a higher Spanish proficiency level by the end of the study abroad, forgot more than participants with a lower recall score at T2 (i.e., the difference between the two lines in the respective subplot in Figure 5.9 increases with an increase in T2 performance score). Finally, participants' own judgment of attrition severity was also predictive of observed forgetting rates: people who thought their Spanish got worse were the ones who indeed forgot the most (Figure 5.9). None of the other participant-level predictors significantly modulated forgetting.

On the item level, we observed a main effect of word frequency, such that successful recall in the Spanish vocabulary test, regardless of when it was administered (T2 vs. T3), was more likely for higher frequency items compared to lower frequency items. The interaction term between frequency and Session did not reach significance, but there was a numerical trend for high frequency items to be forgotten less than low frequency items (p = .060). Cognate status, in turn, significantly predicted forgetting, such that forgetting was more pronounced for non-cognates compared to cognates (see Figure 5.10).

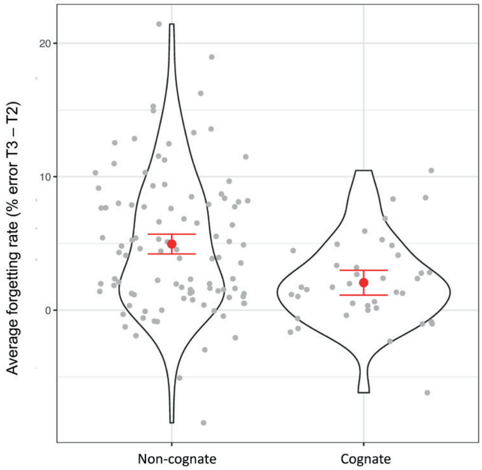

Figure 5.10

Violin plot of forgetting rates (T3 error % - T2 error %) for cognates and non-cognates separately, averaged over participants. Grey dots reflect items, red dots reflect the mean forgetting rate for cognates and non-cognates, respectively. Error bars reflect the standard error around the mean.

5.3.2.2 | *Does Studying Abroad Have Long Term (Linguistic) Benefits?*

From the previous model, we learned that leaving the study abroad destination and returning to one's home country will result in attrition of foreign language vocabulary for the majority of people. A question that arises then is whether a study abroad has any long-term linguistic benefits at all, or whether people return to pre-study-abroad proficiency levels soon after leaving the foreign country. To answer this question, we ran another mixed effects logistic regression with the same random effects structure as the above models with data from all three time points (T1, T2 and T3) and with Session as the only fixed effect (dummy coded with T1 as reference level). The model outcome shows that the study abroad indeed had long-term benefits for our participants. Participants learned while abroad (T2 performance exceeds T1 performance: $\beta = 1.40$, $z = 90.66$, $p = <.001$) and forgot after moving back to Germany (see Figure 5.3). However, since their T2-T3 forgetting rates are smaller than their T1-T2 learning rates, performance at T3 was still significantly better than performance at T1 ($\beta = 1.042$, $z = 68.95$, $p = <.001$).

5.3.2.3 | *Testing the Regression Hypothesis*

A long-standing debate in the attrition literature concerns whether forgetting mirrors acquisition, and hence whether what you learned last is forgotten first. For our study, the Regression Hypothesis, as the former claim is also called, would predict that words learned between T1 and T2 (i.e., words not known at T1, but known at T2) have a higher probability of being forgotten at T3 than words learned before T1 (i.e., words known at both T1 and T2). To test this, we limited our dataset to only the words that were known at T2. For those words, we asked whether their T3 Spanish performance was predicted by their T1 performance. In modelling terms, this corresponds to running a mixed effects logistic regression on T3 Spanish performance with T1 performance as fixed effect (effects coded: -0.5, 0.5). This analysis showed that Spanish words unknown at T1 but known at T2 (learned while abroad) indeed had a higher probability of being forgotten than words that were already known at T1 ($\beta = -1.18$, $z = -30.99$, $p = <.001$; mean accuracy at T3 for words known at T1: 94%, *SD* = 5%; mean accuracy at T3 for words unknown at T1: 73%, *SD* = 15%).

5.3.2.4 | *Does the Type of Spanish Input Matter for Retention Rates?*

Finally, we asked whether the type of input matters for retention. Regardless of the time someone spends speaking Spanish, does someone who receives almost exclusively native input forget less than someone who receives more non-native, and hence potentially faulty Spanish input? This information was obtained for a subset of 47 participants out of the 97 participants, as explained above. We ran the same mixed effects logistic regression model as in the main analysis reported above on this subset with the average amount of native input at T2 and T3 in interaction with Session. The model outcome suggests that the amount of native input someone receives (regardless of the total amount of Spanish input) does predict forgetting rates. Participants who received little to no native input forgot more than people who got a lot of native input, and the latter group in fact appears to show learning rather than forgetting ($\beta = 0.06$, $z = 2.92$, $p = .004$). In order to check whether this finding holds up in a more parsimonious model, we reran the model including the seven significant predictors from the full model in the main analysis above. In this model, shown in Table 5.6, input type still explained a significant amount of the variance in forgetting rates. Moreover, except for cognate status and amount of experience with Spanish (prior to T1), all previously significant predictors were still equally predictive of forgetting in this subset.

Table 5.6

Logistic mixed effects model output for analysis including amount of native input as predictor, run on subset of participants (N = 47).

Fixed effects	Estimate	*SE*	*z*	$p(\chi^2)$
Intercept	**2.21**	**0.18**	**12.21**	**< .001**
Session	**-0.22**	**0.02**	**-9.11**	**< .001**
Participant-level predictors				
Spanish frequency of use	**0.10**	**0.04**	**2.49**	**.013**
Amount of native input	0.03	0.04	0.88	.380
Attrition self-judgment	0.26	0.04	0.67	.504
English letter fluency	0.00	0.04	0.07	.943
German letter fluency	-0.04	0.04	-0.90	.371
Amount of Spanish experience	-0.01	0.04	-0.27	.788
T2 performance	**1.07**	**0.04**	**-0.27**	**< .001**
Session * Spanish frequency of use	**0.21**	**0.03**	**7.09**	**< .001**
Session * Amount of native input	**0.08**	**0.03**	**3.38**	**.001**
Session * Attrition self-judgment	**0.08**	**0.03**	**3.02**	**.003**
Session * English letter fluency	**0.06**	**0.02**	**2.54**	**.011**
Session * German letter fluency	**-0.06**	**0.03**	**-2.32**	**.021**
Session * Amount of Spanish experience	0.03	0.03	1.09	.274
Session * T2 performance	**-0.09**	**0.03**	**-3.50**	**.001**
Item-level predictors				
Cognate status	**1.04**	**0.18**	**5.71**	**< .001**
Session * Cognate status	0.04	0.03	1.34	.182
Random effects	Groups	Var	*SD*	
Item	Intercept	4.57	2.14	
Subject	Intercept	0.05	0.21	

Note. Relevant parts of the table (i.e. interaction terms) are highlighted. Significant effects are marked in bold. *SE* = standard error; $p(\chi^2)$ = Chi-square p-value; Var = variance; *SD* = standard deviation.

5.4 | Discussion

The present study aimed at unravelling the driving forces behind foreign language attrition 'in the wild'. How come we forget foreign language vocabulary, and what determines how fast and severe this lexical forgetting is? Based on recent lab simulations of FL attrition (e.g., Chapter 2), we hypothesized that language use would play a major role in determining the rate of attrition. More specifically, we assumed that continued target language use would positively impact FL retention, and conversely, that use of other languages would negatively influence FL proficiency. We also asked whether it matters which other language an attriter used the most (i.e., L1 vs. L2), and whether we can observe a trade-off in accessibility between those languages and the target foreign language. In a large-scale longitudinal project, we followed a group of German learners of Spanish, who studied abroad in Spain for one semester. We evaluated their Spanish proficiency by means of a picture-naming vocabulary test at the beginning of the study abroad (T1), at the end of it (T2) and roughly six months post return to Germany (T3). Next to the Spanish proficiency test, participants completed fluency tests in German and English at each time point, as well as a questionnaire, asking, among other things, for current frequency of use indications and their motivation to learn Spanish.

By means of a logistic regression model, we then investigated which factors best predicted changes in Spanish vocabulary knowledge from T2 to T3. The terms Spanish proficiency and Spanish retention / attrition rates, as frequently used in the remainder of this manuscript, always refer to *lexical* Spanish proficiency and *lexical* retention / attrition rates, respectively, also when not explicitly stated. The longitudinal design made it possible to obtain fine-grained, participant-specific lexical attrition rates and hence enabled a much more precise analysis of the determinants of individual differences in FL lexical forgetting than would be possible with a cross-sectional study design. Overall, we indeed observed forgetting after the first six months back in Germany: participants performed worse on the Spanish vocabulary test at T3 than at T2. Forgetting rates also varied considerably between individuals. In line with expectations, higher forgetting rates were associated with less frequent use of Spanish during the attrition period. Partially in line with expectations, more forgetting was also associated with relative increases and decreases in letter fluency scores in German and English respectively. In a secondary analysis with a subset of participants, the quality of the continued Spanish input (i.e., during the attrition period) also proved important, with more native input leading to better retention of Spanish.

Because attrition is a complex phenomenon though, we also took variables other than language use into account. The large sample of participants that we tested online allowed for the inclusion of a large number of variables and made it possible to ask not only how predictors modulate forgetting in isolation, but also what their relative contributions to FL attrition are when accounted for together, in one parsimonious model of FL attrition. Out of those additional variables, more years of experience with Spanish prior to the study abroad also predicted better retention rates. Conversely, better performance on the Spanish vocabulary test at the end of the study abroad predicted more forgetting. Finally, neither motivation to learn Spanish nor non-verbal long-term memory capacity or the ratio of German to English use affected Spanish retention rates. These and other findings will each be discussed in detail below.

5.4.1 | The Role of Language Use in Foreign Language Attrition

Despite the undisputed role of language use in theories of language attrition (e.g., Köpke, 2002; Paradis, 2004), we are among the first to establish a clear relationship between target foreign language use and maintenance of FL skills in real attriters. Previous studies with real attriters have paradoxically often failed to observe a consistent relationship between the two (e.g., Bahrick, 1984a,b; Mehotcheva, 2010). As we discussed in the Introduction (section 5.1.2), this failure may stem from the way in which language use was measured in those experiments. Given that many studies are cross-sectional rather than longitudinal in design, participants are often asked to estimate frequency of use once and in retrospect. In the current study, we averaged over multiple measures of frequency of use, taken once every month during the attrition period, asking for current rather than retrospective judgments. Moreover, instead of asking for ratings on a scale or for indications in hours and minutes, we chose for percentages as a measure. Percentages are easier to estimate for participants than absolute hours and because of that they are less prone to differences in subjective perceptions of time. We believe that these aspects combined resulted in a more accurate description of (average) Spanish frequency of use over the attrition period and hence are at least part of the reason why we were able to observe a clear-cut relationship between Spanish frequency of use and Spanish retention. We encourage future research to adopt similarly frequent, percentage-based frequency of use questionnaires.

Admittedly, by using percentages, we ignore information about the total amount of time that someone spends engaging with a language. One might argue that equating people who speak a lot with people who tend to isolate themselves and speak very little is problematic. For our specific population, this turned out not to be of any concern. Additional frequency of use indications in hours (at T2 and T3 only, and separately for a number of different contexts, see Appendix D.7 for a list of questions) showed that our participants did not differ much in the amount of time they spent doing certain activities (all *SD*s < 1.6 hours). We think that this is likely to be true for a lot of foreign language attriter populations, especially when the population is homogenous in terms of age, socio-economic status and cultural background, such as the population we recruited. Our study thus showed that among attriters comparable in terms of total absolute amount of language use, differences in the relative amount of use of different languages are a reliable predictor of forgetting rates.

Next to quantity of Spanish input, we also found that the quality of the input matters. Participants who, regardless of the total amount of time they spent speaking Spanish, received mostly native input forgot less than participants who received less native and hence potentially faultier and less reliable input. In a model with both quantity and quality of exposure (as well as all other predictors), both factors appeared to be equally important. We are not aware of any other study that has explicitly tested whether the amount of native as compared to non-native input matters for foreign language retention. Nevertheless, the finding that the quality of the foreign language input matters resonates well with previous calls to account for the context of language use in studies on (both foreign and first) language attrition (Schmid, 2007, 2019).

Because the amount of Spanish language use is inversely related to the amount of use of all other languages combined (especially when using percentages to elicit language usage patterns), a positive relationship between Spanish retention and use equals a negative relationship between Spanish retention and the use of other languages. In that respect, our findings are in line with lab studies that report that speaking languages other than the target FL language hampers subsequent access to the FL (e.g., Chapter 2; Chapter 3; Levy et al., 2007). It should be noted though that these lab studies' findings are not fully comparable to our results. In the lab, unlike in real life, researchers can keep target language use constant while manipulating non-target language use and can hence investigate the role of speaking other languages on target FL attrition in isolation (see Chapters 2 and 3). While this is impossible in real life, we still asked whether it made a difference which language our participants spoke most during the time they did not speak Spanish. Chapter 2 suggests that other foreign languages, such as English for the participants in the current sample, interfere more with a foreign language than their mother tongue. The English to German ratio variable that we introduced to answer this question did not modulate forgetting rates though, not even in a separate model alone. While this means that we partially failed to replicate the lab findings from Chapter 2, it should be noted that all of our participants were using far more German than English (values are all below one and close to zero). We thus may have not had enough variability to detect the effect that Chapter 2 revealed. For a fair test, one would need to sample sufficient numbers of both participants that use more German than English and participants that use more English than German (or at least enough participants that use both languages equally much, which was also not the case in our sample). In real life, it might be difficult to find participants that use more English than German while living in Germany. An interesting alternative for future research might be to follow Germans that after their study abroad in Spain do not return to Germany, but instead move to a country where they are immersed in English and see whether they suffer more from Spanish attrition than those that return to their L1 environment. For now, we can only conclude that for people who use their L1 a lot more than their L2, it does not matter how much more they use the L1, as L3 attrition rates were comparable regardless of the ratio of German to English use.

Finally, we also took fluency measures in English and German and hypothesized that relative fluency increases in both languages would predict proficiency decreases in Spanish. We found partial evidence for such a trade-off: participants whose German letter fluency scores increased the most relative to other participants, and hence participants who maintained their Germany fluency best (or in fact improved), were indeed more likely to forget Spanish vocabulary than those whose German letter fluency scores did not increase. Conversely though, and unexpectedly, relative increases in English letter fluency predicted increases, rather than decreases, in Spanish proficiency. From our data it thus appears that the two foreign languages co-develop and possibly even facilitate each other. Only the native language shows the trade-off that we had expected based on previous lab research on interference-induced forgetting (e.g., Bailey & Newman, 2018; Chapter 2). While this is puzzling and appears to contradict Chapter 2, it is worth taking a closer look at what the fluency scores reflect. Our trade-off hypothesis was based on the premise that fluency scores can be used as a proxy for verbal ability and ease of access in German and English, and that they would hence correlate positively with frequency of use in German and English respectively. The change in letter fluency from T2 to T3 in German and English, however, did not correlate with average frequency of use between T2 and T3 in German ($r = .03$) and English ($r = .22$), nor did the change in English letter fluency correlate with our participants' perceived change in English proficiency, as assessed via self-ratings ($r = .1$). The same is true for the category fluency scores (all r's < .26). The fluency scores thus do not appear to reflect what we hoped they would. A large part of the variance in fluency scores between participants appears to be caused by factors other than language use and verbal ability. One such factor might be executive control ability, which has often been linked to (especially letter) fluency performance

(Luo et al., 2010; Shao et al., 2014). However, we have no way of knowing whether or to what extent our letter fluency scores reflect differences in executive control ability, or which other factors might be contributing to their variability, making their interaction with Spanish proficiency, and especially the opposite interaction slopes for English and German letter fluency, difficult to interpret. It is also interesting that out of the two types of fluency scores, letter and not semantic category fluency predicted forgetting. Category fluency scores have been much more reliably linked to verbal ability and vocabulary size than letter fluency scores (Shao et al., 2014), and so it would have been more intuitive for changes in category (rather than letter) fluency to correlate with changes in Spanish proficiency. All things considered, the fluency data should be taken with a grain of salt.

For future studies, a test of English and German proficiency (rather than a test of verbal ability) could make for a more straightforward test of the trade-off hypothesis. Should the current pattern prove reliable (i.e., should changes in L3 proficiency indeed correlate positively with changes in L2 and negatively with changes in L1 proficiency), it would mean that interference between two foreign languages is much less prevalent in real life than the lab studies suggest, and that counter to what we concluded in Chapter 2, L1 is the stronger interferer. Maybe English was not used frequently enough by our participants to result in interference for Spanish (English frequency of use was below 15% on average throughout the attrition period, compared to 74% for German), or maybe the interference was present yet so small that it was overwritten by some other, mediating factor. To get a preliminary sense of whether our pattern was reliable or not, we ran a statistical model with the difference in English proficiency self-ratings (T3-T2) instead of English (letter and category) fluency performance. We did not replicate the positive relationship between changes in (perceived) English and (observed) Spanish proficiency. In a model with only the change in English proficiency self-ratings and Session as predictors, we observed a negative relationship, such that participants whose English improved according to their own self-judgments forgot more Spanish than people whose English got worse (β = -0.04, z = -2.24, p = .025). This effect is in line with the trade-off hypothesis. The change in perceived English proficiency, however, no longer significantly predicted forgetting rates when included in the full model with all other predictors (excluding fluency scores, β = -0.01, z = -0.41, p = .683). Though not conclusive, these follow-up analyses cast further doubt on the fluency findings and the suitability of fluency tasks to seek answers to the trade-off hypothesis.

5.4.2 | FL Proficiency Prior to Attrition Onset

Apart from frequency of use, a number of other factors predicted retention rates. Participants who performed better in the vocabulary test at the end of their study abroad, for example, were more likely to forget phonemes from T2 to T3 than participants with poorer T2 performance. This finding appears to contradict earlier research which has shown that a higher level of FL proficiency prior to attrition onset is *beneficial* for foreign language maintenance (e.g., Bahrick, 1984a,b; Mehotcheva, 2010; Murtagh, 2003; Weltens, 1988; but see Engstler, 2012). A closer look at those studies, however, reveals that some of them, in fact, do not assess the effect that prior foreign language proficiency has on forgetting rates (i.e., the change in performance over time), but rather just the effect it has on performance at a single measurement of attrition (i.e., performance at T3 alone rather than the change in performance from T2 to T3; e.g., Mehotcheva, 2010; Murtagh, 2003). That higher initial proficiency predicts better performance at a later testing point is not that surprising and is in fact also the case in our data (see Figure 5.9, lower middle panel, the dashed, pink line has a positive slope), but this does not say anything about the amount of forgetting (i.e., change in knowledge) since the start of the attrition period. A relatively recent longitudinal study that did assess the effect of initial proficiency on the change in performance from a pre-attrition to an attrition measurement, similar to how we do it here, found no evidence for an effect of initial proficiency on forgetting severity (Engstler, 2012).

Another difference between our study and previous research is how we defined 'initial proficiency'. Most previous studies used proficiency self-ratings (Engstler, 2012; Murtagh, 2003; Mehotcheva, 2010), the number and level of courses taken in the FL prior to attrition onset, and past course grades (or a combination thereof; Bahrick, 1984a,b; Weltens, 1988) as estimates of initial proficiency. Our T2 performance measure is much more specific than that: it is an objective measure of past FL vocabulary knowledge and given that it reflects prior performance on the same task, it is directly comparable to our attrition (T3) measurement. To the best of our knowledge, we are the first to ask specifically how pre-attrition vocabulary size influences lexical forgetting rates.

Nevertheless, it might seem puzzling that we observe a *negative* rather than a positive effect of T2 performance on subsequent Spanish vocabulary retention rates. That someone who knew more words is more likely to forget parts (i.e., phonemes) of those words over time, however, might just be a reflection of the fact that they had more to lose in the first place. Knowing more words might mean that they knew some of those words less well and hence were more likely to forget them subsequently. One might wonder then whether participants who knew more at T2 forgot more only in absolute terms (the number of forgotten phonemes, as our model suggests), or also in *relative* terms (the percentage of forgotten words out of all words known at T2 baseline); absolutely more is not necessarily relatively more, it might in fact even be relatively less. To answer this specific question, we ran a linear model on participants' relative forgetting rates ((T3-T2)/T2)) with all the participant-level predictors from the main mixed model (thus excluding item-level predictors and random effects per participant or item, which are not possible in this simplified model).[15] Partially in line with what Bahrick (1984a,b) and Weltens (1988) reported[16], this model revealed that participants with high T2 scores forgot *less* in relative terms compared to people with lower T2 scores. In summary, we thus find that high T2 performers forget *more words/ phonemes in absolute terms*, yet *fewer words/phonemes in relative terms* (i.e., a smaller percentage of their original knowledge) than low T2 performers, which means that the retention rate of the former group was ultimately better.

5.4.3 | Amount of Experience with the Foreign Language

Next to T2 performance, the amount of experience with Spanish prior to the study abroad also predicted forgetting rates in the main analysis. In line with previous research, participants with more years of Spanish experience were less likely to forget. In the Introduction, we discussed how including amount of FL experience might appear redundant given that more Spanish experience prior to the study abroad is likely to at least partly contribute to better FL proficiency at the end of

[15] We chose the generalized mixed model for the main analysis, because the GLMM makes use of the full dataset (repeated measures per participant), rather than requiring aggregate data. In the linear model on relative forgetting rates, the data is reduced to just one value per participant and a lot of valuable information gets lost, which comes at the cost of accuracy and sensitivity in estimating fixed effects. For the interpretation of fixed effects in our main model, it furthermore needs to be remembered that the estimate of any fixed effect (a single predictor or an interaction term) in our main model reflects the effect that this predictor has when all other predictors in the model are held constant (at their respective average for our scaled continuous predictors). For the effect that Spanish frequency of use has on forgetting rates (i.e., the Frequency*Session interaction), for example, our model shows that all other things being equal, someone who uses more Spanish forgets less. With T2 performance being equal, forgetting absolutely less also corresponds to forgetting relatively less (a smaller proportion). For the assessment of all predictors other than T2 performance itself, our main model thus already implicitly takes initial T2 performance into account. Predictor plots based on relative forgetting rates (as used in the additional linear model described here) confirm this and can be inspected in Appendix D.12.

[16] Bahrick (1984a,b,) and Weltens (1988) found that participants with different levels of initial training forget the *same amount in absolute terms*, which however corresponds to less forgetting in relative terms (a smaller proportion of their original knowledge). While we also find less forgetting in relative terms, our results differ from these previous studies in that we report *more* forgetting rather than comparable forgetting in *absolute* terms.

the study abroad and hence should partially reflect the same as our T2 performance measure. The two variables, however, only mildly correlated with one another in our sample (r = .21), and consequently predicted different aspects of the variance in forgetting rates. As already discussed in the Introduction, it is very possible that some participants are faster learners than others and hence reach the same level of proficiency in less time. What is more, years of exposure is not necessarily indicative of the quality or even quantity of exposure to the language in those years. Interestingly, in an additional analysis with the amount of native language input during the attrition period as extra predictor, the amount of experience with Spanish prior to the study abroad was no longer predictive of forgetting, while T2 performance and frequency of use of Spanish during the attrition period continued to explain a large part of the variance. This pattern suggests, quite intuitively, that amount and quality of *recent* language use are more important for successful retention than total amount of time spent learning a foreign language. This pattern emerged from a model with only 47 out of the total 97 participants though, and hence needs replication before firm conclusions can be drawn (remember that we only had data regarding the quality of input from 47 participants and hence ran the secondary model on this smaller subset).

Finally, it should be noted that unlike Mehotcheva (2010), we did not use the length of the study abroad as a predictor in the final model. Study abroad length was significantly worse at explaining forgetting rates than prior amount of experience with Spanish. This is not surprising given that there was little variability in the length of the study abroad between our participants and the fact that participants differed much more in terms of how many years they had been learning Spanish before the study abroad. Not including this variable in the final model was furthermore reinforced by Mehotcheva's (2010) failure to observe a relationship between study abroad length and Spanish retention rates.

5.4.4 | Motivation

Based on Gardner's Attitude and Motivation test battery, we asked participants about their integrative motivation (i.e., intrinsic desire to learn Spanish for social reasons), as well as their instrumental motivation (i.e., desire to learn Spanish for practical reasons) and their anxiety to speak Spanish. In separate models, all three motivational variables significantly modulated forgetting, such that higher integrative and instrumental motivation and lower anxiety to speak Spanish were beneficial for Spanish retention. Out of those three scores, even though integrative motivation might intuitively seem most relevant, only instrumental motivation entered the final model. Including all three scores was not possible with the current sample size, and instrumental motivation scores predicted forgetting rates slightly better than the other two in those separate models.

Though highly predictive of forgetting rates when included in a model on its own, instrumental motivation no longer predicted forgetting rates once frequency of use and other factors were also accounted for. The most likely reason for this is the relatively high correlation between Spanish frequency of use and instrumental motivation (r = .49). Participants who were highly motivated to learn and maintain Spanish likely sought out more opportunities to speak it. By virtue of being correlated, the two variables explain partially the same variance; frequency of use, however, appears to do so better. This pattern demonstrates once again that FL attrition is a complex phenomenon and that many of the variables that are thought to contribute to it interact with one another in complex ways (see also Mehotcheva & Mytara, 2019), and that consequently, as many variables as possible need to be considered together in one parsimonious model in order not to overestimate the contribution of single predictors.

That we do not observe a beneficial effect of motivation in our main model is in line with previous failures to establish such a link consistently (e.g., Mehotcheva, 2010; Xu, 2010; though see Wang, 2010). As noted in the Introduction, previous studies often only used one measure of motivation, acquired at the attrition time point. In the present study, we administered the motivation questionnaire at each session, thus enabling us to capture potential changes in motivation from before attrition onset to the attrition measure. Interestingly, on average, neither instrumental, nor integrative motivation or anxiety to speak Spanish changed much from T2 to T3. There were considerable individual differences though, highlighting the necessity of collecting more than just one motivation measure. By calculating the *average* over these two scores, we were able to take this variability into account. As explained in the methods section, we reasoned that this average would provide the most accurate approximation of each participant's *overall average* motivation to learn Spanish throughout the attrition period.

5.4.5 | Non-Verbal Memory Capacity

Non-verbal long-term memory capacity did not predict forgetting either, neither in the full model nor in a separate model by itself. To our knowledge we were the first to assess the relationship between non-verbal and verbal long-term memory in a population of foreign language attriters and our results suggest that the two do not interact and that ability in one domain does not predict ability in the other. Based on just the current results, however, this conclusion is premature. The Doors test that we administered is a test of non-verbal, episodic recognition memory. That the memories that are encoded in the Doors test are episodic rather than semantic in nature (like foreign language vocabulary after consolidation) and the fact that the test ended in a receptive rather than a productive recall test might explain why the Doors test had such poor explanatory power in the model on productive Spanish vocabulary forgetting. Future studies might want to consider a test of productive non-verbal memory instead, which within the scope of this online experiment, however, was unfortunately not possible to implement. Another possibility is that participants did not pay enough attention during the encoding phase and that their performance on the Doors test thus does not reflect their true non-verbal long-term memory capacity. Even though participants performed above the chance level of 25% on average, performance overall was still rather low (M = 50%). The online set-up and the fact that the encoding phase did not require responses from participants and was automatically paced possibly encouraged participants to take the test less seriously. Future research should improve upon those aspects. Either way, a more complete assessment of non-verbal long-term memory capacity is necessary before drawing firm conclusions on its relation to verbal memory capacity.

5.4.6 | Attrition Self-Judgments

Participants who thought their Spanish decreased indeed actually forgot more than participants who thought their Spanish had not changed (or had even improved). Participants who thought they improved, however, on average still performed worse on the Spanish vocabulary test at T3 than T2, suggesting that, unlike in other studies before (e.g., Weltens, 1988), our participants did not overestimate their attrition, but rather underestimated how much they forgot. This underestimation might be related to the fact that participants were asked to judge their proficiency overall, rather than their vocabulary knowledge specifically. Given that some of the words we asked for in the vocabulary test were low-frequency words and given that the attrition self-judgment was provided prior to the vocabulary test, our participants might not have been aware of their vocabulary gaps and hence rated their overall proficiency decline as less severe than the vocabulary test suggests.

5.4.7 | Item Level Predictors: Cognate Status and Word Frequency

Finally, on the item level, we partially confirmed Weltens' (1988) and de Groot and Keijzer's (2000) findings. In the main statistical model with all participants and predictors included, cognates were retained better than non-cognates. In a subsequent model with only a subset of the participants and with a slightly different fixed effects structure (i.e., the model including quality of input as predictor, for which we only had data for 47 participants), this pattern was no longer evident. It is unclear what exactly caused this difference, though we speculate that it is related to power issues with the smaller sample size in the secondary analysis. It thus remains unclear whether cognates indeed are superior to non-cognates in memory (also see Engstler, 2012, for a failure to find a beneficial effect of cognate status on retention rates in returnees from abroad).

For word frequency, we also only found partial evidence for a modulating effect. In a separate, initial model with only Spanish word frequency, low frequency words were more likely to be forgotten than high frequency words. This effect, however, was no longer statistically robust ($p = .06$) in the final model with all other predictors combined. Our study thus suggests that word frequency plays only a minor role in vocabulary retention.

5.4.8 | Does Forgetting Follow the Reverse Order of Acquisition?

Next to investigating individual differences in foreign language attrition, we also asked whether we could find evidence for regression in foreign language attrition. In its original formulation, the Regression Hypothesis (RH) states that forgetting follows the reverse order of acquisition (Jakobson, 1941) and hence that recently learned words are forgotten faster than words learned long ago (i.e., remotely). The longitudinal design of our study and the fact that we have a pre-study abroad baseline in addition to the pre-attrition baseline provided a unique opportunity to test this. For our participants, this would mean that words they learned during their study abroad (known at T2 but not T1) should be more likely to be forgotten than words they already knew before the study abroad (known at both T1 and T2). Our data indeed suggest that this is the case and hence provide evidence for regression in foreign language lexical attrition.

Some researchers have proposed that it is not the information learned last, but rather the information learned least well that is forgotten first (Hedgcock, 1991). With the current dataset, we have no way of testing whether it is order or degree of learning that matters and hence cannot rule out the alternative hypothesis. Words that were recently learned, however, are likely to also be learned less well, because they have had less chance to consolidate than remotely learned words. It might hence be difficult to disentangle the two hypotheses in practice. Future research might want to take a look at reaction times in picture naming as an estimate of how well the words were known at T2. If degree of learning matters more, naming latencies at T2 should be a better predictor of T3 performance than order of acquisition. The audio quality of the Spanish recordings and the insufficient control over presentation timing online unfortunately made reaction time analyses impossible for the current study.

5.4.9 | A Final Note on Overall Attrition Rates and Study Design

Regardless of the individual differences discussed so far, it is worth noting that observing significant attrition effects after just six months is quite remarkable, especially in light of the fact that some previous studies failed to observe any attrition after much longer periods of disuse (e.g., Engstler, 2012; Murtagh, 2003; Weltens, 1988). That we do observe forgetting in this rather short attrition period is probably in part due to the fact that we tested productive vocabulary recall, rather than receptive FL skills, as was the case in Weltens (1988). Much to the reassurance of all language learners out there though, forgetting after the study abroad was overall still rather small (4% on average) and much less pronounced than learning during the study abroad (17%), meaning that the study abroad did ultimately result in long-term linguistic gains for our participants despite ensuing attrition.

Long-term in this case, of course, refers to only six months. We know from Bahrick's (1984a,b) research that FL language skills decline steadily for the first three to six years, suggesting that the 4% loss that we observed on average after six months is just the beginning of the attrition process, and that we would have likely observed much larger forgetting rates had we tested our participants a few years after the study abroad. From the range of forgetting scores that we observed (-12% to 21%), it also becomes clear that not everyone forgot: some in fact continued learning (16 out of 97). Participants who still used Spanish regularly, with friends in Spain or at university as part of their studies, were more likely to improve from T2 to T3 rather than to attrite. What we call the 'attrition period' throughout this paper is hence only an attrition period in the aggregate. What is more, some of our participants, and in particular those that were tested later and hence had a somewhat longer 'attrition period' (time between T2 and T3), returned to Spain for their summer vacation. The longer the T2-T3 interval in our study, the higher the chance for re-exposure to Spanish, which is opposite to what attrition length usually should denote. We therefore refrained from including this variable in our statistical analyses. Amount of exposure to Spanish was already (and much better so) captured by our regular frequency of use questionnaires.

As a final note, we would like to briefly discuss the overall set-up of the experiment and its advantages and disadvantages. We explicitly chose a longitudinal design because testing the same participants at multiple time points allows for accurate, participant-specific forgetting rates. Cross-sectional designs instead rely on a nonattriting control group that needs to be comparable in all other aspects to the groups of attriters that are assessed. Given that language attrition is such a highly individual and complex phenomenon, finding such a control group is difficult. Longitudinal designs thus have a distinct advantage over cross-sectional approaches, and are to be preferred, whenever possible, in analyses of individual differences in (FL) attrition.

When testing participants in person, following a large enough group of attriters over a long period of time is very time-consuming. Online testing offers a convenient way out. Participant drop-out rates will still be high, possibly even higher than in real life, but the automaticity of online testing and the fact that one is no longer constrained to one geographical location make recruitment much easier and hence enable the researcher to test much bigger samples than would be possible with in-person testing. Naturally, online testing comes with its own set of issues that need to be carefully considered before embarking on such a study. Most prominently these issues include technical difficulties (e.g., unstable internet connection, compatibility issues with outdated browsers, poor audio equipment on participants' laptops) and constraints on the type of tasks that can be administered online (e.g., free speech data will be difficult to elicit online), but also the lack of control over how seriously participants engage in the tasks. Some of those issues can be overcome with thorough piloting of the software and are likely to become less and less problematic as online testing tools advance and become publicly available, especially with the current surge in online research in times of Covid-19. Overall, we thus recommend the online deployment of experiments for the study of foreign language attrition and we hope the current study highlights the virtues of the approach. Most importantly, we think that such online testing opens up opportunities for truly large-scale longitudinal studies at a relatively low (administrative) cost to the researcher and that these benefits clearly outweigh the costs outlined above.

5.4.10 | Conclusion

The present study investigated individual differences in foreign language attrition as it unfolds within the first months after an immersive study abroad. In a longitudinal fashion, we followed German university students throughout their study abroad in Spain, as well as throughout their first roughly six months back in Germany. The longitudinal set-up enabled the precise calculation of individual forgetting rates of vocabulary knowledge. In line with expectations, yet counter to most previous research on foreign language attrition 'in the wild', variation in forgetting rates, to a large part, turned out to be due to differences in the quantity and quality of language use: more frequent Spanish language use was clearly beneficial for Spanish vocabulary retention, and so was native (as compared to non-native) Spanish input, regardless of the total amount of input they received. Conversely, and partially in support of recent lab-based simulations of FL attrition, increases in L1 German but not L2 English verbal ability appeared to be detrimental for FL retention, at least in an environment where L1 use is predominant and where there is little room for use of other languages. Next to language use, more experience with Spanish prior to the study abroad was also partially beneficial for Spanish retention. Furthermore, Spanish vocabulary knowledge prior to attrition onset also affected retention rates: participants who knew more words on average at T2 forgot absolutely more, yet in relative terms they forgot less than participants who knew fewer words at T2. Motivation to learn the foreign language and non-verbal memory capacity, in turn, had no influence on Spanish maintenance. Overall, the present study thus provided empirical evidence for the importance of continued language use for FL maintenance in real attriters. Moreover, using a longitudinal design and state-of-the-art statistical analyses, we were able to shed light on the complex interplay between language use and other determinants of FL attrition.

CHAPTER 6

General Discussion

Parts of this chapter are based on:

Mickan, A., McQueen, J. M., & Lemhöfer, K. (2019). Bridging the gap between second language acquisition research and memory science: the case of foreign language attrition. *Frontiers in Human Neuroscience, 13*, 397. https://doi.org/10.3389/fnhum.2019.00397

6.1 | General Discussion

The experiments reported on in this thesis were designed to further our understanding of why we forget (words from) foreign languages. Is it time alone that drives forgetting, or are there other processes that contribute to it? I approached this question from two angles. Inspired by the memory literature on forgetting, Chapters 2 to 4 studied foreign language (FL) attrition in the lab by evaluating the conditions that do and those that do not successfully induce it. Chapter 5, in turn, took a more traditional approach and observed the phenomenon in 'the wild' in a group of natural attriters. Below, I will first summarize the main findings, and then discuss what we have learned from these studies overall and in what way the lab approach complements the more traditional way of studying FL attrition.

6.1.1 | Summary of Findings

In the domain-general memory literature, it has been proposed that forgetting can be the consequence of interference and competition from related memories (Anderson, 2003). Inspired by research on such interference-induced forgetting, **Chapter 2** asked whether similar dynamics are at the basis of FL (lexical) attrition. If so, FL vocabulary forgetting should be inducible through the more recent use of the same words in other languages. In line with interference theory, **Chapter 2** showed that this was indeed the case. Retrieval practice in both L1 Dutch and L2 English hampered subsequent recall of recently learned L3 Spanish words. Participants were slower and less accurate at recalling Spanish words when they had previously retrieved their L1 or L2 translation equivalents compared to when they had not. These interference effects were not just momentary and transient, but in fact persisted long-term (20 minutes in accuracy, one week in reaction times), making between-language competition and interference plausible mechanisms for real-life foreign language attrition. What is more, in reaction times, the interference effect was (at least after 20 minutes) more pronounced for the English compared to the Dutch interference group. Retrieval practice in another foreign language thus appears to be somewhat more detrimental for later L3 recall ability than interference from one's mother tongue. The second experiment reported on in this chapter suggests that this is most likely due to frequency of use differences between native and non-native languages: less frequently used languages (or low frequency words within a language) are harder to retrieve and because of that induce stronger interference than frequently used languages (or words). Overall, **Chapter 2** thus showed that interference from the recent use of other languages is indeed a driving force in FL attrition and that this interference is especially prominent if it comes from less frequently used languages.

Chapter 3 provides corroborating neural evidence in favor of the idea that accessibility difficulties in a foreign language can be the direct consequence of the more recent use of other languages. We replicated the main effects from **Chapter 2** with L3 Italian as the to-be-forgotten foreign language and L2 English as (the only) interfering language: participants were slower and less accurate to recall L3 Italian words for which they had retrieved L2 English translations compared to words for which they had not. In the EEG, these interference effects were accompanied by an increased N2, more power in the theta frequency band and a decreased LPC. The former two signatures have been linked to interference and competition processes in other cognitive domains and hence are in line with the idea that the behavioral forgetting effects we observe at final test are the result of competition and interference between translation equivalents. The LPC, in turn, appears to index the consequences of this interference, namely reduced accessibility to the interfered Italian labels compared to the not interfered ones. What is more, we were able to link activity during the interference phase in English to retrieval speed at final test in Italian. Words that took participants long to recall in Italian at final test, and that had hence been most strongly interfered with, showed an increased N2 during their previous retrieval in English. Retrieval difficulty thus does not only emerge at final test, but is instead already set in motion during the preceding interference phase.

Chapters 2 and 3 together showed that the mere retrieval of words from other languages can induce forgetting of recently learned foreign language words. In **Chapter 4**, we showed that the opposite is also true: learning new (Spanish) FL words can hamper subsequent access to the same words in an already well-known, other foreign language (English). The negative after-effects of L3 Spanish learning were visible in both L2 English retrieval speed and accuracy, and they emerged immediately after (or possibly during) learning and did not grow stronger with time. The latter tentatively suggests, unlike we had hypothesized, that the newly learned L3 Spanish words do not need to be fully integrated into the mental lexicon through offline consolidation to interact and interfere with their L2 English counterparts. As we discuss in more detail in **Chapter 4** though, more research on the role of consolidation in new-learning-induced FL attrition is necessary to confirm this conclusion.

Finally, **Chapter 5** documents the role of language use for FL attrition in a population of natural attriters. Following a group of German native speakers throughout a study abroad in Spain as well as during their first six months back in Germany, we observed a clear link between continued L3 Spanish use and L3 Spanish retention. German native speakers who no longer used Spanish regularly when back in Germany showed the most severe forgetting rates, while people who still used Spanish forgot less, or even continued to learn. Interestingly, it did not matter which other language our participants spoke the most during the time they did not speak Spanish: those who spoke only L1 German forgot just as much or little as those who used both L2 English and L1 German in equal amounts. On top of that, we also report partial evidence in favor of between-language interference in the wild: participants whose German fluency scores increased the least while back in Germany incurred the smallest losses in Spanish vocabulary knowledge. Next to the quantity of language use, the quality also mattered: participants with a higher proportion of native (as compared to non-native) Spanish input, forgot less. Finally, our data showed that FL attrition is a multifaceted phenomenon that is not only the consequence of current language use patterns. More experience with the foreign language before going abroad was also beneficial for Spanish vocabulary retention and participants who knew more words at the end of the study abroad forgot relatively less (though more in absolute terms) than participants with a smaller pre-attrition vocabulary size.

6.1.2 | The Role of Language Use and Between-Language Competition for FL Attrition

Together, the studies in this thesis show how language use impacts the maintenance and conversely the forgetting of foreign language skills. While Chapter 5 illustrates the beneficial role of FL use for FL vocabulary retention and a tentative trade-off between FL and L1 verbal ability, Chapters 2 to 4 consistently show that both the use and the new learning of words in other languages can lead to retrieval difficulties for the same words in an either recently learned or an already well-known FL. Even though linking language use to language attrition might seem trivial, previous research had often failed to do so (see Mehotcheva & Mytara, 2019, and Chapter 5 for a discussion of possible reasons). The research in this thesis thus adds important empirical evidence to the debate of how language use shapes linguistic knowledge, both positively and negatively, and in doing so, advances our understanding of why we forget foreign languages.

The majority of previous studies on FL attrition were observational in nature, meaning that they either compared different groups of natural attriters to one another and to a group of learners of the FL (cross-sectional studies; e.g., Bahrick, 1984a,b; Hansen & Chen, 2001), or followed a set of attriters over a certain period of time (longitudinal studies; e.g., Murtagh, 2003; Tomiyama, 2008). The latter design, though most suitable to the study of individual differences in FL attrition, is very time-consuming and often participants drop out along the way. Because of that, the majority of *large-scale* studies on FL attrition to date are cross-sectional studies. Chapter 5, instead, is one of few large-scale, *longitudinal* studies on the topic. Having multiple measures of FL ability over time enabled us to calculate participant-specific forgetting rates and thus made it possible to arrive at a much more fine-grained measure of attrition than many previous studies on the topic. On top of that, we administered monthly, percentage-based frequency of use questionnaires to get a more continuous and reliable estimate of frequency of use throughout the attrition period. Thanks at least in part to these design improvements, we are among the first to convincingly document the beneficial role of FL use for FL maintenance.

The lab studies in Chapters 2 to 4, in turn, take a different, less conventional approach to test the relevance of language use for FL attrition. Inspired by research on interference-induced forgetting in the domain-general memory literature, we tried to simulate attrition in the lab under tightly controlled experimental conditions. So-called retrieval-induced forgetting (RIF) studies had previously established that the repeated retrieval of category-exemplar pairs (e.g., FRUIT – banana) interferes with the subsequent retrieval of unpracticed exemplars from the same category (e.g., FRUIT – strawberry; Anderson et al., 1994). We reasoned that if similar dynamics underlie FL attrition, it should be possible to induce FL vocabulary forgetting by having participants retrieve translation equivalents. Like category-exemplar pairs, translation equivalents are related to one another by virtue of being connected to the same concept (rather than the same semantic category) and hence should compete with one another when cued with that concept (i.e., in picture naming). Chapters 2 and 3 demonstrated that this parallel is justified and that between-language competition can indeed induce FL attrition in the lab: retrieving the English label for 'dog' made it difficult and in some cases entirely prevented subsequent retrieval of the Spanish label for dog ('perro'). Similarly, retroactive interference studies in the memory literature had shown that the learning of new associations (A-C) hampers later retrieval of previously learned associations (A-B; e.g., Barnes & Underwood, 1959). Chapter 4 extended those findings to the language domain by showing that learning the Spanish label 'perro' for the picture of a dog complicates later retrieval of its English label, despite the fact that the English label was learned long ago and was well known to our participants.

Competition and inhibition lie at the heart of these interference effects. In Chapters 2 and 3, we reasoned that L2 English (and L1 Dutch) retrieval during the interference phase is hindered by competition from the concurrent activation of the recently learned L3 Spanish/Italian words. Subsequent inhibition of these competing L3 labels (or alternatively boosting of the L2 English / L1 Dutch labels) leads to the observed competition disadvantage for these L3 words at final test (compared to L3 words who did not have to be inhibited because their L1/L2 translation equivalents were not intermittently retrieved). This competition disadvantage (i.e., the interference effect) is larger when the interference phase took place in L2, because L2 (like low frequency L1) words experience relatively more competition from L3 words during interference than L1 (or high frequency) words and hence require more inhibition of the competing L3 words for their own successful retrieval (see Chapter 2 for details). Likewise, in Chapter 4, we speculated that the L2 English words are difficult to retrieve at final test because they experience competition from the recently learned L3 Spanish words and/or possibly because they were previously inhibited to facilitate the learning of the L3 Spanish words. Together, the results from Chapters 2 to 4 clearly and consistently establish between-language competition as a plausible mechanism behind foreign language forgetting. The interference account of forgetting from the domain-general memory literature thus appears to be applicable to FL attrition.

6.1.2.1 | *Interference vs. Facilitation Between Translation Equivalents*

In light of some of the findings in the bilingual speech production literature, it might be surprising that our experiments revealed such robust inhibitory effects between translation equivalents. Studies with bilingual picture-word interference (PWI) paradigms, for example, have sometimes reported the opposite, namely facilitated naming of a picture in L2 or L1 if its translation equivalent is presented alongside (e.g., Costa et al., 1999; Dylman & Barry, 2018; Hermans, 2004; see Chapter 1). Similarly, some blocked language switching studies have reported a facilitative effect of L1 naming on subsequent naming of the same pictures in L2 (e.g., Branzi et al., 2014; Wodniecka et al., 2020; see Chapter 3). These observations stand in stark contrast to the interference effects that we and other language RIF studies report (e.g., Bailey & Newman, 2018; Isurin & McDonald, 2001; Levy et al., 2007) and beg the question how these conflicting results can be reconciled.

By definition, and as already explained earlier, translation equivalents reflect two possibilities of referring to one single concept. On the conceptual level, they are thus identical and can facilitate each other's processing; on the word form level, in turn, they are in conflict with one another and can hence interfere with each other's retrieval. Whether facilitation or interference prevails ultimately appears to depend on the specific experimental design that is used to elicit the effect. The PWI facilitation effects, for example, are only observed when the distractor is presented visually and when it appears on screen *before* the to-be-named picture. When the distractors appear too late (i.e., after the to-be-named picture), they no longer facilitate naming significantly (e.g., Costa et al., 1999) and when they are presented auditorily they even interfere with rather than facilitate retrieval of their translation equivalents (e.g., Melinger, 2018). Moreover, when the L1 distractor is phonologically related (rather than identical) to the L1 translation equivalent, L2 picture naming is slowed down rather than sped-up through the presentation of this so-called phono-translation distractor (e.g., Hermans et al., 1989).

Likewise, in the blocked language switching studies, the beneficial effect of having named in L1 prior to naming in L2 is observed only relative to a 'no prior naming' baseline (i.e., L2 naming after no previous naming at all). The visual and conceptual familiarity with the already named picture then simply overrides the simultaneously unfolding inhibitory effects on the word form level (see Chapter 3 for details). One could have expected similar facilitation effects in our experiments. Given that the concepts of the interfered items were more often and more recently accessed, they might have been primed and easier to retrieve at final test, despite the fact that naming was required in a different language. Our design, however, with the initial learning phase in Chapters 2 and 3 (or the pre-test in Chapter 4), minimized the effect that additional retrieval attempts during the interference phase could have on conceptual availability and enabled us to observe the interference effects instead. Unlike some researchers have claimed, our results thus show that translation equivalents *do* compete with one another and hence support language non-selective models of bilingual lexical access that assume that words in both languages are co-activated and compete with one another for selection.

6.1.3 | The Benefits of Inducing Foreign Language Attrition in the Lab

Though language competition has long been a part of theoretical models of (FL) attrition (e.g., ATH, Paradis, 1993, 2004), support for it as an underlying mechanism in FL attrition previously mostly came from observations of code-switches and language intrusions or syntactic transfer errors in natural attriters. Observing these outcomes of the attrition process, however, does not implicate between-language competition as a causal mechanism behind them. Simulating FL attrition rather than observing it in real life, instead, allows for such causal inferences. Manipulating the presence or absence of a presumed cause of forgetting (e.g., interference) while keeping all other potentially relevant factors (e.g., degree of learning, target language use) constant, enables the researcher to determine which conditions do and which do not lead to forgetting. In showing that attrition can be induced through the controlled retrieval of other languages, the experiments reported in Chapters 2, 3 and 4 are thus among the first to directly test the assumptions underlying Paradis' ATH and show that the use of other languages can indeed be (at least) one of the causes of FL forgetting (see also Bailey & Newman, 2018; Levy et al., 2007; Isurin & McDonald, 2001 and Chapter 2 for a detailed comparison of these studies with our experiments).

Chapter 5 highlights another, though related advantage of the lab approach. In real life, the amount and frequency of target foreign language use is inversely related to the use of all other languages a person speaks (L1 and other FLs): you only have a fixed amount of time, and using most of that time speaking one language naturally leaves less time for speaking other languages. Because of that, a positive relationship between target FL use and FL maintenance automatically entails a negative relationship between FL maintenance and the use of all other languages combined. In showing that Spanish use has a positive effect on Spanish retention, Chapter 5 thus also provides evidence in line with the idea that non-Spanish (i.e., English and German) language use is detrimental for Spanish retention. This support, however, is only indirect. In fact, the inverse relationship between target and non-target language use in real life, makes it impossible to isolate the role of non-target language use alone in natural attriters. The lab approach, in turn, allows to do precisely that. Because it enabled us to selectively manipulate the presence and absence of (for example) English language use, while keeping Spanish use (and other aspects) constant, the lab approach is unaffected by the otherwise inverse relationship between the two.

The ability to tease apart aspects of language processing which in real life are inherently confounded makes the lab approach very useful. Enabling almost full control over not only the interference phase but also the initial FL learning situation, the lab approach has the potential to help answer long-standing open questions regarding FL attrition. The Regression Hypothesis (RH; also discussed in Chapter 5) is a case in point. In its original formulation, the RH posits that we tend to forget first the information (e.g., words) we learned last (Jakobson, 1941). It has been claimed, however, that the order of acquisition itself might not actually matter as much as the degree of learning of a given word (i.e., 'best learned = last forgotten'; Hedgcock, 1991). In the real world, these two theories are almost impossible to tease apart: with more time for rehearsal and repetition, remotely learned words will be better encoded than recently learned words. To worsen matters further, the first words one learns in a new language tend to be the most frequent; later learned words or structures instead are usually less frequent, harder to learn, and possibly more vulnerable to forgetting because of their difficulty rather than order of acquisition. A lab study could disentangle these options by manipulating the acquisition order during the initial learning phase while keeping the amount of exposure (and thus the degree of learning) for each word – as well as subsequent interference – equal.

In a similar fashion, lab paradigms could test whether active use of other languages (as in Chapters 2-4) is necessary to induce forgetting, or whether mere passive exposure to other languages is enough. Evidence from memory studies seems to suggest that active retrieval and response generation (though not necessarily successful retrieval, Hellerstedt and Johansson, 2016; Storm et al., 2006) is necessary to induce RIF; passive exposure or even reading out loud of exemplar-pairs does not induce forgetting of related items (Anderson et al., 2000; Bäuml, 2002). Restricting or controlling participants' language use is impossible in real life, and so this question would again be challenging to address in an observational study with real attriters. Finding that passive exposure to a language does not lead to forgetting would have interesting implications for educational strategies though. As these two examples show, the lab paradigms can be adjusted in multiple ways (both their learning and the interference phase) and hence have the potential to investigate a great number of different questions. I hope that the research in this thesis encourages future studies to consider approaching FL attrition in a similar way.

6.1.4 | To What Extent Do the Lab Studies Model Real-Life FL Attrition?

While the previous section outlined the virtues of the lab approach, it also needs to be acknowledged that the lab studies simplify the FL attrition situation and break it down to a level that is not intuitively representative of real life. As with any lab study, the experimental control one gains comes at the cost of ecological validity. The level of proficiency that we simulated in Chapters 2 and 3, for example, is much lower than in most real-life attrition scenarios. Likewise, the amount of interference we induce does not scale up to the extent of interference attriters are likely to experience in real life. Although this simplification is exactly what makes the lab approach so useful (see above advantages), it is worth discussing in some more detail to what extent the studies in Chapters 2 to 4 capture and model real life and in how far the forgetting effects we observe resemble forgetting 'in the wild' and what that means for the conclusions we can draw from them. To do so, I will first compare and discuss the lab studies among themselves and ultimately end in a comparison of the lab results with the natural attrition results from Chapter 5.

6.1.4.1 | *The Time Frame of Forgetting Effects in the Lab*

First of all, one might wonder how comparable the time frame of the lab studies is to real life attrition. Arguably, for the lab-induced interference effects to be a plausible mechanism for real-life foreign language attrition, they need to persist long-term. Previous studies on between-language competition in bilingual speech production have often restricted their search and analysis of interference effects to single trials, that is to the interfering effect of co-activation *during* online language processing rather than its after-effects (though see Kleinman & Gollan, 2018). Chapters 2 and 3 showed that the detrimental effect of retrieving words in another language persists for at least 20 minutes (a common delay in studies on long-term memory, e.g., Anderson et al., 1994). Beyond that, Chapter 2 additionally showed that interference effects can persist for up to an entire week and hence that between-language interference has true long-term ramifications.

Admittedly, a delay of one week is still rather minuscule compared to the time frames of months or even years that are typically studied in observational attrition studies. In experimental terms, however, it is quite remarkable for effects to persist for an entire week. While looking at longer time delays would be theoretically interesting for future studies, doing so only makes sense if (1) it can be guaranteed that the participants are not re-exposed to the target language within that time, and (2) only if additional interference can be reliably quantified. If these two conditions are not met, the experimenter would no longer have the experimental control that makes the simulation approach so useful. What is more, it would be difficult to interpret the outcome of a longer time delay. Additional interference through the intermittent use of other languages would happen equally often for items in the interference and no interference conditions, and so would wash the interference effect out. The experimentally induced interference effect might thus disappear with time, however, not necessarily because it is not long-lasting, but instead because of additional interference, the very mechanism that caused the effect in the first place. There is thus a logical limit to the length of the delays one can sensibly look at while maintaining experimental control; and one week is arguably already stretching this limit.

On this note, it should also be mentioned that even when interference effects persist long-term, this persistence does not entail that the memory in question has been lost entirely or permanently. Regardless of the delay, and in simulation studies and observational studies alike, a vocabulary test at a given moment in time can only test temporary (in)accessibility and hence can only establish whether (and how easily) information is *currently* available to conscious memory retrieval. Whether a memory is truly lost (i.e., 'forgotten' in laymen's terms) is impossible to test. Chapter 2, in fact, shows that the forgetting we induced was definitely temporary for a substantial number of items: about a third of the words that our participants in Experiment 1 were unable to recall 20 minutes after interference were successfully recovered and recalled a week later. From studies on relearning though, we already know that most forgetting in real life is temporary anyway (e.g., de Bot et al., 2004; Ebbinghaus, 1885, 1913; see Chapter 1). That our experiments only induce temporary 'forgetting' thus does not make them less representative of forgetting in real life. The often temporary nature of forgetting is key to definitions of forgetting in both the memory and language attrition literature, where forgetting is typically described as temporary retrieval failure (e.g., Ebbinghaus, 1885, 1913; Roediger et al., 2010) or -in linguistic terms- as a performance problem characterized by accessibility difficulties rather than structural loss (e.g., Sharwood Smith, 1989).

6.1.4.2 | *Quantifying Forgetting*

Relatedly, for all of our lab experiments, we used both reaction times and accuracy as indicators of forgetting rates. Retrieval failure, as measured in recall accuracy, is the most straightforward, and perhaps the most convincing evidence of (at least temporary) forgetting. Reaction times, in turn, might seem like an unusual choice to quantify the phenomenon and one might wonder to what extent they really are indicative of forgetting. Our decision to include reaction times was in part based on tradition in psycholinguistics, where they are often the measurement of choice, not least for quantifying online between-language competition effects. In the memory domain, reaction times are much less typical, especially in experiments on interference-based forgetting, even though the experimental study of forgetting actually started with time-sensitive measurements as well (see Ebbinghaus, 1913, who took the time to relearn as an indication for a memory's strength). In Chapter 2, we argued that reaction times (i.e., naming latencies in our experiments) are just as important and relevant for measuring forgetting as recall accuracy. We reasoned that forgetting is a gradual process that initially manifests in increased retrieval difficulty (i.e., slowed down retrieval) and that only eventually ends in (temporary) retrieval failure. By that definition, prolonged naming latencies are the natural precursor to forgetting. Naming latencies are also a much more fine-grained measure of recall ease than accuracy. In our studies, in fact, they revealed interference-related forgetting effects that otherwise would have gone unnoticed: for example, the language difference in Chapter 2, or the interference effects in Chapter 4 in general (which were not present in accuracy in Exp.1, but were robust in RTs). We thus encourage future research on (FL) attrition to likewise supplement naming accuracy measures with naming latencies.

6.1.4.3 | *The Magnitude of Interference in and Across Lab Studies*

While the lab studies draw a consistent picture and repeatedly establish the effect of interference on subsequent FL recall ability, it appears that these effects differ slightly in magnitude between chapters. Looking specifically at the interference effects in accuracy rates, they appear to be strongest when the to-be-forgotten language is weakest. We report the most robust effects in Chapter 2 where L3 Spanish words were learned in just one session and had only one night to consolidate before interference was induced. Effects were less pronounced in Chapter 3 where the learning session was spread over two days and where Italian words had more time to stabilize before the English interference phase. Finally, in Chapter 4, where the to-be-interfered English words were learned long-ago rather than in the context of the experiment itself, we report the least reliable accuracy effects. Do these differences in interference magnitude suggest that interference is less relevant in real life, where foreign language knowledge is usually well known before it is forgotten?

First of all, given that we *do* observe interference effects in both naming latencies and accuracy in *all* chapters, even Chapter 4, it cannot be concluded that interference is entirely irrelevant in situations where the foreign language is well consolidated (see Chapter 4 for a discussion). What is, however, possible is that better learned material requires *more* interference events to be affected. Participants in Chapter 3, for example, underwent a longer learning session (mean of 15 exposures per item spread over two days) than those in Chapter 2 (mean of 12 exposures in one day), yet received the same amount (and type) of interference (9 retrieval attempts per item in English). It would be interesting for future research to test whether increasing the number of interference trials makes up for the longer learning session and hence whether the ratio of degree of learning (i.e., N exposures during learning) and extent of interference (i.e., N retrieval attempts during interference) matters for the magnitude of forgetting. If so, this would mean that interference as a general mechanism is effective regardless of how well-established the foreign language knowledge is and that forgetting can always be induced provided there is a sufficient amount of interference.

Next to degree of learning and extent of interference, the type of interference is also likely to play a role. Hence, another reason why the interference effects were small in Chapter 4 might be that interference consisted of new learning rather than retrieval of already known material. Although we did not observe a significant increase in interference with consolidation of the newly learned Spanish words, it might still be the case that the newly learned Spanish words needed more time and possibly repeated exposure to fully interfere with their English translations (i.e., to approximate the effect that the well-known interferers in Chapter 2 had). Again, future research will be necessary to fully understand the impact that the type of interference has on interference severity.

Finally, and on an entirely different note, the interference effect might be smallest in Chapter 4 because the experimental design afforded the biggest chance for conceptual priming (i.e., facilitation) effects to counteract the inhibitory effects. As discussed earlier, translation equivalents can both facilitate and interfere with one another. In Chapters 2 and 3, words in both the interference and no interference conditions were initially primed through the long learning sessions; so much so that a few additional retrievals during the interference phase likely did not result in significant additional conceptual priming for the interfered items[17], making it possible to instead observe clear interference effects on the word level. In Chapter 4, in contrast, there was no initial learning phase, the pre-test consisted of only one retrieval attempt. The items that were subsequently learned in Spanish (interference condition) are thus likely to have benefitted conceptually from the extra retrievals while words in the no interference condition whose concepts were not accessed and primed any further did not. The conceptual facilitation advantage of interfered over not interfered items was then likely bigger in Chapter 4 than Chapters 2 and 3, and the interference effects might thus have been smaller in Chapter 4 simply because they were washed out more.

On the basis of the current set of experiments, it is impossible to pinpoint the interference magnitude differences to one single aspect. Ultimately, interference success or strength might depend on a combination of all of the above discussed aspects: the degree of learning of the to-be-forgotten FL words, the type and extent of interference induction and the experimental set-up used to test the effect. Until we know more about the relationship between these aspects, it would thus be premature to conclude that interference (as a mechanism in general) is less applicable to well-consolidated foreign language material, or that the lab paradigms are ill-suited for the study of real-life FL attrition because of that.

6.1.4.4 | *The Use of Pictures as Stimuli in Lab Simulations of FL Attrition*

Another, final aspect of the lab studies that merits discussion is the use of pictures, and more specifically the use of the *same* pictures for all experimental sessions (learning, interference and final test). The EEG data reported in Chapter 3 pointed towards one problem associated with the use of identical pictures for both interference and final test: the recent exposure to pictures during the interference phase made the items in the interference condition easier to process and recognize in the first round of the final test (reduced N400). Next to contaminating the EEG signal, this visual recognition facilitation (and possibly conceptual facilitation on top of that, see previous section) could have worked against the interference effects we were looking for. The fact that the interference effect was stronger than this

[17] We know from previous research that the benefit of each additional picture naming trial on response latencies decreases rapidly with every retrieval (e.g., Gollan et al., 2005; Griffin & Bock, 1998).

picture repetition (and the conceptual priming) effect actually serves to reinforce the robustness of the interference effect. Nevertheless, future studies might want to avoid the picture repetition confound by using different sets of pictures during interference and final test.

Alternatively, one could argue that using the same pictures enhanced the interference effects, the argument here being that participants build up stimulus-response mappings rather than concept-word form mappings during the learning phases of our experiments and that the interference effects we observe are dependent on the specific stimuli (i.e., pictures) we used (see also Anderson, 2003, for a related discussion on cue (in)dependency in RIF studies). It could be argued that the interference phases of our experiments disrupt this stimulus-response mapping through the association of a new response to the same stimulus (i.e., the same picture). In Chapters 2 and 3, for example, the interference phase tasks added an English verbal response to the picture cue and hence made retrieval of the first association (the Spanish / Italian response) difficult. Likewise, in Chapter 4, the initial mapping between the picture and the English response might have been overwritten through the lengthy learning session in Spanish which led to the predominant association of the picture cue with the Spanish verbal response. According to this stimulus-response mapping account, using the same pictures might not only have facilitated the interference effects, it might in fact be the only reason why we observed interference at all.

While it is possible that our effects are boosted through the use of one set of pictures, I believe that the interference effects are unlikely to stem solely from this design aspect. First of all, the interactions of interference with language and frequency in Chapter 2 would be difficult to explain from a pure stimulus-response mapping point of view. If our interference effects were the result of only the disruption of an arbitrary stimulus-response mapping, it should not matter for the magnitude of interference what kind of additional (i.e., interfering) vocal response (e.g., L1 vs. L2) is paired with a given picture. What is more, if the interference we observed was purely a reflection of stimulus-response mapping dynamics, you would expect the interference effects to be strongest in Chapter 4, where the interfering stimulus-response mapping is most reinforced (through a Spanish learning session with on average 13 exposures compared to 8 retrievals in the English/Dutch interference phase in Chapters 2 and 3) and where the target response-mapping is least well established (in just one retrieval in the English pre-test in Chapter 4 compared to an average of 12 and 15 exposures in the learning sessions in Chapters 2 and 3 respectively). As discussed in the previous section though, this was not the case and the interference effects were in fact smallest in Chapter 4. The stimulus-response mapping account thus cannot accommodate the differences in interference magnitude within and between our experiments. Next to these theoretical considerations, it also needs to be kept in mind that using multiple sets of pictures requires careful matching of these pictures in terms of visual complexity and similarity as well as naming agreement, which is not trivial. Nevertheless, considering both shortcomings (recognition facilitation and possible stimulus-response specificity), I encourage future studies to take on the challenge and to verify that the effects persist when memory for the interfered and not interfered words is probed with an independent set of pictures at final test.

6.1.4.5 | *Reconciling Contradictory Results from Chapter 5 with Those in Chapter 2*

Finally, returning to the comparability of the lab studies to real life, it should be discussed that the results from Chapter 5 do not *fully* align with the results from Chapter 2 (which comes closest to modelling the attrition scenario in Chapter 5). We only found partial evidence for a trade-off between English and German fluency measures on the one hand and Spanish vocabulary knowledge on the other. Moreover, we also found no evidence for the fact that L2 English interferes more than L1 German. Do the subtleties of the interference effects revealed in the lab studies therefore not apply to real-world FL attrition? As we discuss in detail in Chapter 5, more research will be necessary to answer that question, but solely on the basis of Chapter 5, such a conclusion would be premature.

First of all, the fluency measures were not the most ideal choice to investigate the trade-off hypothesis. We had collected the fluency measures to estimate how easily our participants could access words in their L1 and L2 and had hoped to observe a trade-off in accessibility such that an increase in L1 / L2 fluency would go hand in hand with a decrease in accessibility to Spanish vocabulary, as indexed in an overall performance decrease. It turned out, however, that the fluency measures did not, as we had hoped for, capture verbal ability alone and thus cannot be used as a direct proxy for ease of accessibility in a given language. Possibly, the online set-up added to this issue: some participants produced very few words in the fluency tests and it is unclear whether this was because they could not access more words, or because they did not try hard enough (and did not feel encouraged enough to do so online). With regard to the second inconsistency, our participant sample was very homogenous in terms of language use and did not cover the full spectrum of English to German use, making it difficult (if not impossible) to observe the language differences from Chapter 2. With these limitations in mind, it seems more likely that the findings from Chapter 5 fail to capture the nuances of interference rather than that the lab studies reveal entirely irrelevant interference patterns.

6.1.5 | A Call for an Integrated, Complementary Approach to FL Attrition

As the above sections reveal, there is clearly need for more research into interference-based foreign language forgetting, both more lab research and more large-scale longitudinal studies with real attriters. With respect to observational, longitudinal research, I hope that Chapter 5 has highlighted the benefits (and drawbacks) of online testing, as well as the necessity to administer frequent, timely frequency of use and proficiency measures whenever possible. For the lab approach, I encourage future research to follow up on the studies presented in this thesis. Replicating, for example, the secondary findings in Chapters 2 and 4 or the main effects with a second set of pictures would help settle some of the open question discussed above. Otherwise, the possibilities using variations of the lab paradigms presented here are manifold, as I hope the above examples illustrate. Overall, I hope to have shown that using paradigms from the memory literature complements more traditional approaches to attrition and offers a fresh look at the phenomenon which allows for the investigation of questions that would otherwise remain unanswered. Both approaches have their limitations and the lab approach is by no means meant to replace traditional, observational studies on FL attrition though. Instead, I believe the field would benefit from a healthy balance between studies using the two approaches.

6.1.6 | Going Beyond Lexical FL Attrition

6.1.6.1 | *Foreign Language Syntactic Attrition*

The work in this thesis focuses on foreign language *lexical* attrition. Yet, speaking a language requires much more than just mastery of its words. Similarly, attrition is by no means limited to the forgetting of vocabulary; grammatical structures also attrite (e.g., Hansen, 1999; Tomiyama, 2008). It would be interesting to extend the current line of research to cover these other types of attrition as well. As a first step, one could look at grammatical gender, for which negative transfer and interference effects are well documented in online processing (e.g., Lemhöfer et al., 2008). One might ask whether retrieving L1 gender for a set of nouns (e.g., 'der Strand', the beach, masculine in German) interferes with and makes people forget just recently learned, but incompatible FL gender assignments (e.g., 'het strand', the beach, neuter in Dutch). For more rule-governed aspects of grammar, the design might need some adjustments. The learning phase, for example, will most likely need to be longer, and include tasks other than just picture naming for participants to learn FL grammatical rules. Moreover, the control (i.e., no-interference) condition will need to be carefully chosen. If the syntactic property is not item specific, one would need to find a syntactic rule that is comparable in complexity, yet not in conflict with and thus not prone to interference from the L1 (or another FL). This might prove challenging for some aspects of grammar and some language combinations. In such cases, one might need to resort to between-subject designs and compare a group that learns a conflicting rule with a group that learns the same rule but in a language that implements this rule similarly to their L1. Though somewhat more challenging, extending the approach pioneered in Chapters 2 - 4 to syntax would be a very interesting line of research.

6.1.6.2 | *First Language Attrition*

Finally, one might also wonder to what extent the lab approach advocated here is applicable to L1 attrition. Much like Chapter 4, a study on L1 attrition would not involve an initial learning session. Instead, one would start with a baseline L1 picture naming test. This baseline speed and accuracy measurement would then be followed by an interference phase that could consist of learning of some of the same words in a new FL, or (more akin to Chapters 2 and 3 and as in Levy et al., 2007) could consist of retrieval tasks for some of the same words in an already known foreign language. Finally, retrieval speed and accuracy would be measured again for all words in L1. Clearly, such a study is perfectly conceivable and has, in fact, been conducted by Levy et al. (2007). It is unclear though how easily deeply engrained L1 knowledge can be interfered with in the lab. Even though Levy et al. (2007) observed worse L1 English recall rates after L2 Spanish retrieval practice, Runnqvist and Costa (2012) were later unable to replicate this finding. Levy et al. (2007) used a rather indirect measure of recall ability (rhyme-generation rather than picture naming), which possibly underestimated L1 productive knowledge and was heavily influenced by aspects other than L1 lexical retrieval ease. Moreover, Chapter 4 suggests that FL words learned long ago might be slightly more difficult to interfere with and this might be even more true for L1 words. The interference phase might need to be longer or spaced out over multiple days for successful L1 attrition induction. Regardless though, it should be mentioned that the L1 words under investigation (just as the FL words in Chapter 4) will have been learned in the wild and not under controlled circumstances. Hence, there will be limitations to the types of simulations one can run; disentangling the effects of order of acquisition vs. degree of learning of L1 words on L1 attrition rates (i.e., the RH, see earlier), for example, would be impossible to address.

6.1.7 | Conclusion

Foreign language attrition is a frustrating yet common phenomenon among multilinguals. In this thesis, I asked why it is that we forget words from foreign languages. I approached this question from two angles: on the one hand through tightly controlled experiments that aimed at inducing vocabulary forgetting in the lab, and on the other hand in a longitudinal study with real attriters. The latter approach is reflective of how FL attrition is typically investigated. Chapter 5, however, also showed how the traditional, observational approach to FL attrition can be improved upon and is among the first experiments to establish a clear-cut role of language use for FL maintenance. Chapters 2, 3 and 4, in turn, illustrate that using paradigms from the memory literature complements the traditional approach in important ways. By testing which conditions lead to forgetting and which do not, they allow for causal rather than just correlational inferences and thereby advance our understanding of the processes underlying foreign language attrition. Using this approach, Chapters 2, 3 and 4 provided convincing and converging evidence that between-language competition is at least one of the driving mechanisms behind foreign language attrition. Naturally, each approach comes with its own set of advantages and disadvantages, some of which I have discussed in detail in this last Chapter. Overall, though, I hope to have shown that the lab approach is a promising avenue for future investigations into (FL) attrition and that a sound mixture of both approaches- as adopted in this thesis- is crucial if we are to understand what it means to forget a foreign language.

References

Abbasian, R., & Khajavi, Y. (2010). Lexical attrition of general and special English words after years of non-exposure: The case of Iranian teachers. *English Language Teaching, 3*(3), 47–53.

Agren, T. (2014). Human reconsolidation: A reactivation and update. *Brain Research Bulletin, 105*, 70–82. https://doi.org/10.1016/j.brainresbull.2013.12.010

Alharthi, T., & Al Fraidan, A. (2016). Language use and lexical attrition: Do they change over time? *British Journal of English Linguistics, 4*(1), 50–63.

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language, 49*(4), 415–445. https://doi.org/10.1016/j.jml.2003.08.006

Anderson, M. C. (2015). Incidental forgetting. In A. Baddeley, M. W. Eysenck, & M. C. Anderson (Eds.), *Memory* (2nd ed., pp. 231–263). Psychology Press.

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. https://doi.org/10.1037/0278-7393.20.5.1063

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review, 7*(3), 522–530. https://doi.org/10.3758/BF03214366

Antony, J. W., Ferreira, C. S., Norman, K. A., & Wimber, M. (2017). Retrieval as a fast route to memory consolidation. *Trends in Cognitive Sciences, 21*(8), 573–576. https://doi.org/10.1016/j.tics.2017.05.001

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* [Database]. Retrieved from http://celex.mpi.nl/

Baddeley, A. D., Emslie, H., & Nimmo-Smith, I. (1994). *The Doors and People Test: A test of visual and verbal recall and recognition*. Thames Valley Test Company.

Baddeley, A. D., Hitch, G. J., Quinlan, P. T., Bowes, L., & Stone, R. (2016). Doors for memory: A searchable database. *Quarterly Journal of Experimental Psychology, 69*(11), 2111–2118. https://doi.org/10.1080/17470218.2015.1087582

Baddeley, Alan D. (2015). What is memory. In A. Baddeley, M. W. Eysenck, & M. C. Anderson (Eds.), *Memory* (2nd ed., pp. 3–20). Psychology Press.

Bahrick, H. P. (1984a). Fifty years of second language attrition: Implications for programmatic research. *The Modern Language Journal, 68*(2), 105–118.

Bahrick, H. P. (1984b). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology : General, 113*(1), 1–29. https://doi.org/10.1037/0096-3445.113.1.1

Bahrick, H. P., & Phelphs, E. (1987). Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(2), 344–349. https://doi.org/10.1037/0278-7393.13.2.344

Bailes, C., Caldwell, M., Wamsley, E. J., & Tucker, M. A. (2020). Does sleep protect memories against interference? A failure to replicate. *PLoS ONE, 15*(2), e0220419. https://doi.org/10.1371/journal.pone.0220419

Bailey, L., & Newman, A. J. (2018). Retrieval-induced forgetting and second language vocabulary acquisition: Insights from a Welsh-language training study. In J. M. Fawcett (Chair) (Ed.). *Control Processes in Human Memory: The Role of Retrieval Suppression and Retrieval Practice*. Symposium conducted at the joint meeting of the Canadian Society for Brain, Behaviour and Cognitive Science and the Experimental Psychology Society, St. John's, NL, Canada.

Bakker, I., Takashima, A., van Hell, J. G., Janzen, G., & McQueen, J. M. (2014). Competition from unseen or unheard novel words: Lexical consolidation across modalities. *Journal of Memory and Language, 73*(1), 116–130. https://doi.org/10.1016/j.jml.2014.03.002

Bakker, I., Takashima, A., van Hell, J. G., Janzen, G., & McQueen, J. M. (2015a). Changes in theta and beta oscillations as signatures of novel word consolidation. *Journal of Cognitive Neuroscience, 27*(7), 1286–1297. https://doi.org/10.1162/jocn_a_00801

Bakker, I., Takashima, A., van Hell, J. G., Janzen, G., & McQueen, J. M. (2015b). Tracking lexical consolidation with ERPs: Lexical and semantic-priming effects on N400 and LPC responses to newly-learned words. *Neuropsychologia, 79*, 33–41. https://doi.org/10.1016/j.neuropsychologia.2015.10.020

Bardel, C., & Falk, Y. (2007). The role of the second language in third language acquisition: The case of Germanic syntax. *Second Language Research, 23*(4), 459–484. https://doi.org/10.1177/0267658307080557

Barkat-Defradas, M., Gayfraud, F., Köpke, B., & Lefebvre, L. (2019). Linguistic regression in bilingual patients with Alzheimer's disease. In M. S. Schmid & B. Köpke (Eds.), *The Oxford handbook of language attrition* (pp. 136–145). Oxford University Press.

Barnes, J. M., & Underwood, B. J. (1959). "Fate" of first-list associations in transfer theory. *Journal of Experimental Psychology, 58*(2), 97–105. https://doi.org/10.1037/h0047507

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

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

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology, 80*(3), 1-46. https://doi.org/10.1037/h0027577

Bäuml, K.-H. (2002). Semantic generation can cause episodic forgetting. *Psychological Science, 13*(4), 356-360. https://doi.org/10.1111/1467-9280.00464

Bentin, S., & McCarthy, G. (1994). The effects of immediate stimulus repetition on reaction time and event-related potentials in tasks of different complexity. Journal of Experimental Psychology. *Learning, Memory, and Cognition, 20*(1), 130–149. https://doi.org/10.1037/0278-7393.20.1.130

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–345.

Borovsky, A., Elman, J. L., & Kutas, M. (2012). Once is enough: N400 indexes semantic integration of novel word meanings from a single exposure in context. *Language Learning and Development, 8*(3), 278–302. https://doi.org/10.1080/15475441.2011.614893

Bowers, J. S., Davis, C. J., & Hanley, D. A. (2005). Interfering neighbours: The impact of novel word learning on the identification of visually similar words. *Cognition, 97*(3), B45–B54. https://doi.org/10.1016/j.cognition.2005.02.002

Branzi, F. M., Martin, C. D., Abutalebi, J., & Costa, A. (2014). The after-effects of bilingual language production. *Neuropsychologia, 52*(1), 102–116. https://doi.org/10.1016/j.neuropsychologia.2013.09.022

Brehm, L., Jackson, C. N., & Miller, K. L. (2019). Speaker-specific processing of anomalous utterances. Quarterly *Journal of Experimental Psychology, 72*(4), 764–778. https://doi.org/10.1177/1747021818765547

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The bank of standardized stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS ONE, 5*(5), e10773. https://doi.org/10.1371/journal.pone.0010773

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology, 58*, 412–424. https://doi.org/10.1027/1618-3169/a000123

Choi, J., Broersma, M., & Cutler, A. (2017). Early phonology revealed by international adoptees' birth language retention. *Proceedings of the National Academy of Sciences, 114*(28), 7307–7312. https://doi.org/10.1073/pnas.1706405114

Christoffels, I., Firk, C., & Schiller, N. O. (2007). Bilingual language control: An event-related brain potention study. *Brain Research, 1147*, 192–208. https://doi.org/10.1016/j.brainres.2007.01.137

Christoffels, I., Kroll, J. F., & Bajo, T. (2013). Introduction to bilingualism and cognitive control. *Frontiers in Psychology, 4*, 199. https://doi.org/10.3389/fpsyg.2013.00199

Cohen, A. D. (1989). Attrition in the productive lexicon of two Portuguese third language speakers. *Studies in Second Language Acquisition, 11*(2), 135–149. https://doi.org/10.1017/S0272263100000577

Colomé, À. (2001). Lexical activation in bilinguals' speech production: Language-specific or language-independent? *Journal of Memory and Language, 45*(4), 721–736. https://doi.org/10.1006/jmla.2001.2793

Costa, A., & Caramazza, A. (1999). Is lexical selection in bilingual speech production language-specific? Further evidence from Spanish–English and English–Spanish bilinguals. Bilingualism: *Language and Cognition, 2*(3), 231–244. https://doi.org/10.1017/S1366728999000334

Costa, A., Caramazza, A., & Sebastián-Gallés, N. (2000). The cognate facilitation effect: Implications for models of lexical access. *Journal of Experimental Psychology: Learning Memory and Cognition, 26*(5), 1283–1296. https://doi.org/10.1037//0278-7393.26.5.1283

Costa, A., Miozzo, M., & Caramazza, A. (1999). Lexical selection in bilinguals: Do words in the bilingual's two lexicons compete for selection? *Journal of Memory and Language, 41*(3), 365–397. https://doi.org/10.1006/jmla.1999.2651

Costa, A., & Santesteban, M. (2004). Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners. *Journal of Memory and Language, 50*(4), 491–511. https://doi.org/10.1016/j.jml.2004.02.002

Costa, A., Santesteban, M., & Ivanova, I. (2006). How do highly proficient bilinguals control their lexicalization process? Inhibitory and language-specific selection mechanisms are both functional. *Journal of Experimental Psychology: Learning Memory and Cognition, 32*(5), 1057–1074. https://doi.org/10.1037/0278-7393.32.5.1057

Coutanche, M. N., & Thompson-Schill, S. L. (2014). Fast mapping rapidly integrates information into existing memory networks. *Journal of Experimental Psychology: General, 143*(6), 2296–2303. https://doi.org/10.1037/xge0000020

Crawley, M. J. (2007). *The R book*. John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470515075

Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica, 32*(2), 133–143.

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 364*(1536), 3773–3800. https://doi.org/10.1098/rstb.2009.0111

de Bot, K., & Clyne, M. (1989). Language reversion revisited. *Studies in Second Language Acquisition, 11*(2), 167–177. https://doi.org/10.1017/S0272263100000590

de Bot, K., Martens, V., & Stoessel, S. (2004). Finding residual lexical knowledge: The "Savings" approach to testing vocabulary. *International Journal of Bilingualism, 8*(3), 373–382. https://doi.org/10.1177/13670069040080031101

de Bot, K., & Weltens, B. (1995). Foreign language attrition. *Annual Review of Applied Linguistics, 15*, 151–164. https://doi.org/10.1017/S026719050000266X

de Groot, A. M., & 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. https://doi.org/10.1111/0023-8333.00110 de Vos, J. F., Schriefers, H., & Lemhöfer, K. (2018). Noticing vocabulary holes aids incidental second language word learning: An experimental study. *Bilingualism: Language and Cognition, 22*(3), 500–515. https://doi.org/10.1017/S1366728918000019

Declerck, M., & Philipp, A. M. (2017). Is there lemma-based language control? The influence of language practice and language-specific item practice on asymmetrical switch costs. *Language, Cognition and Neuroscience, 32*(4), 488–493. https://doi.org/10.1080/23273798.2016.1250928

Dewaele, J. M. (1998). Lexical inventions: French interlanguage as L2 versus L3. *Applied Linguistics, 19*(4), 471–490. https://doi.org/10.1093/applin/19.4.471

Dugas, L. G. (1999). *Attrition of pronunciation accuracy among advanced American learners of French.* [Unpublished doctoral dissertation]. Indiana University.

Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science, 18*(1), 35–39. https://doi.org/10.1111/j.1467-9280.2007.01845.x

Dylman, A. S., & Barry, C. (2018). When having two names facilitates lexical selection: Similar results in the picture-word task from translation distractors in bilinguals and synonym distractors in monolinguals. *Cognition, 171*, 151–171. https://doi.org/10.1016/j.cognition.2017.09.014

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology* (H.A.Ruger & C.E.Bussenius, Trans.; original work published 1885). Teachers College, Columbia University.

Ecke, P. (2004). Language attrition and theories of forgetting: A cross-disciplinary review. *International Journal of Bilingualism, 8*(3), 321–354. https://doi.org/10.1177/13670069040080030901

Ellenbogen, J. M., Hulbert, J. C., Stickgold, R., Dinges, D. F., & Thompson-Schill, S. L. (2006). Interfering with theories of sleep and memory: Sleep, declarative memory, and associative interference. *Current Biology, 16*(13), 1290–1294. https://doi.org/10.1016/j.cub.2006.05.024

Engstler, C. (2012). *Language retention and improvement after a study abroad experience* [Unpublished doctoral dissertation, Northwestern University]. https://www.linguistics.northwestern.edu/documents/dissertations/linguistics-research-graduate-dissertations-engstlerdissertation2012.pdf

European Commission (2020). *Erasmus+ annual report 2018*. https://doi.org/10.2766/989852

Ferreira, C. S., Marful, A., Staudigl, T., Bajo, M. T., & Hanslmayr, S. (2014). Medial prefrontal theta oscillations track the time course of interference during selective memory retrieval. *Journal of Cognitive Neuroscience, 26*(4), 777–791. https://doi.org/10.1162/jocn_a_00523

Finkbeiner, M., Almeida, J., Janssen, N., & Caramazza, A. (2006). Lexical selection in bilingual speech production does not involve language suppression. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 32*(5), 1075–1089. https://doi.org/10.1037/0278-7393.32.5.1075

Finnigan, S., Humphreys, M. S., Dennis, S., & Geffen, G. (2002). ERP 'old/new' effects: Memory strength and decisional factor(s). *Neuropsychologia, 40*(13), 2288–2304. https://doi.org/10.1016/S0028-3932(02)00113-6

Folstein, J. R., & Petten, C. V. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: A review. *Psychophysiology, 45*(1), 152–170. https://doi.org/10.1111/j.1469-8986.2007.00602.x

Forcato, C., Argibay, P. F., Pedreira, M. E., & Maldonado, H. (2009). Human reconsolidation does not always occur when a memory is retrieved: The relevance of the reminder structure. *Neurobiology of Learning and Memory, 91*(1), 50–57. https://doi.org/10.1016/j.nlm.2008.09.011

Ford, R. M., Keating, S., & Patel, R. (2004). Retrieval-induced forgetting: A developmental study. *British Journal of Developmental Psychology, 22*(4), 585–603. https://doi.org/10.1348/0261510042378272

Garcia-Bajos, E., Migueles, M., & Anderson, M. C. (2009). Script knowledge modulates retrieval-induced forgetting for eyewitness events. *Memory, 17*(1), 92–103. https://doi.org/10.1080/09658210802572454

Gardner, R. C. (1985). *The Attitude and Motivation Test Battery Manual.* University of Western Ontario. http://publish.uwo.ca/~gardner/

Gardner, R. C., Lalonde, R. N., Moorcroft, R., & Evers, F. T. (1987). Second language attrition: The role of motivation and use. *Journal of Language and Social Psychology, 6*(1), 29–47. https://doi.org/10.1177/0261927X8700600102

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition, 89*(2), 105–132. https://doi.org/10.1016/S0010-0277(03)00070-2

Gollan, T. H., Forster, K. I., & Frost, R. (1997). Translation priming with different scripts: Masked priming with cognates and noncognates in Hebrew-English bilinguals. *Journal of Experimental Psychology: Learning Memory and Cognition, 23*(5), 1122–1139. https://doi.org/10.1037//0278-7393.23.5.1122

Gollan, T. H., Montoya, R. I., Fennema-Notestine, C., & Morris, S. K. (2005). Bilingualism affects picture naming but not picture classification. *Memory and Cognition, 33*(7), 1220–1234. https://doi.org/10.3758/BF03193224

Gollan, T. H., Montoya, R. I., & Werner, G. A. (2002). Semantic and letter fluency in Spanish-English bilinguals. *Neuropsychology, 16*(4), 562–576. https://doi.org/10.1037/0894-4105.16.4.562

Gollan, T. H., & Silverberg, N. B. (2001). Tip-of-the-tongue states in Hebrew–English bilinguals. *Bilingualism: Language and Cognition, 4*(1), 63–83. https://doi.org/10.1017/S136672890100013X

Gómez-Ariza, C. J., Lechuga, M. T., & Pelegrina, S. (2005). Retrieval-induced forgetting in recall and recognition of thematically related and unrelated sentences. *Memory & Cognition, 33*(8), 1431–1441. https://doi.org/10.3758/BF03193376

Goral, M. (2004). First-language decline in healthy aging: Implications for attrition in bilingualism. *Journal of Neurolinguistics, 17*(1), 31–52. https://doi.org/10.1016/S0911-6044(03)00052-6

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition, 1*(2), 67–81. https://doi.org/10.1017/S1366728998000133

Grendel, M. (1993). *Verlies en herstel van lexicale kennis* [Unpublished doctoral dissertation, Radboud University Nijmegen]. https://hdl.handle.net/2066/120134

Griffin, Z. M., & Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language, 38*(3), 313–338. https://doi.org/10.1006/jmla.1997.2547

Gruber, T., Malinowski, P., & Müller, M. M. (2004). Modulation of oscillatory brain activity and evoked potentials in a repetition priming task in the human EEG. *European Journal of Neuroscience, 19*(4), 1073–1082. https://doi.org/10.1111/j.0953-816X.2004.03176.x

Gürel, A. (2004). Selectivity in L2-induced L1 attrition: A psycholinguistic account. *Journal of Neurolinguistics, 17*(1), 53–78. https://doi.org/10.1016/S0911-6044(03)00054-X

Hammarberg, B. (2001). Roles of L1 and L2 in L3 production and acquisition. In J. Cenoz, B. Hufeisen, & U. Jessner (Eds.), *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives* (pp. 21–41). Multilingual Matters.

Hansen, L. (1999). Not a total loss: The attrition of Japanese negation over three decades. In L. Hansen (Ed.), *Second language attrition in Japanese contexts* (pp. 142–153). Oxford University Press.

Hansen, L. (2001). Language attrition: The fate of the start. *Annual Review of Applied Linguistics, 21*, 60–73. https://doi.org/10.1017/S0267190501000046

Hansen, L., & Chen, Y.-L. (2001). What counts in the acqusition and attrition of numeral classifiers? *Japanese Association for Language Teaching Journal, 23*(1), 90–110.

Hansen, L., & Newbold, J. (2001). Literacy as an anchor for the spoken language: Evidence from adult attriters of L2 Japanese. In P. Robinson, M. Sawyer, & S. Ross (Eds.), *Second language acquisition research in Japan* (pp. 101–109). Japan Association for Language Teaching.

Hansen, L., Umeda, Y., & McKinney, M. (2002). Savings in the relearning of second language vocabulary. *Language Learning, 52*(4), 653–678. https://doi.org/10.1111/1467-9922.00200

Hanslmayr, S., Pastötter, B., Bäuml, K.-H. T., Gruber, S., Wimber, M., & Klimesch, W. (2008). The electrophysiological dynamics of interference during the Stroop task. *Journal of Cognitive Neuroscience, 20*(2), 215–225. https://doi.org/10.1162/jocn.2008.20020

Hanslmayr, S., Staudigl, T., Aslan, A., & Bäuml, K.-H. T. (2010). Theta oscillations predict the detrimental effects of memory retrieval. *Cognitive, Affective and Behavioral Neuroscience, 10*(3), 329–338. https://doi.org/10.3758/CABN.10.3.329

Hardwicke, T. E., Taqi, M., & Shanks, D. R. (2016). Postretrieval new learning does not reliably induce human memory updating via reconsolidation. *Proceedings of the National Academy of Sciences, 113*(19), 5206–5211. https://doi.org/10.1073/pnas.1601440113

Hedgcock, J. (1991). Foreign language retention and attrition: A study of regression models. *Foreign Language Annals, 24*(1), 43–55. https://doi.org/10.1111/j.1944-9720.1991.tb00440.x

Hellerstedt, R., & Johansson, M. (2014). Electrophysiological correlates of competitor activation predict retrieval-induced forgetting. *Cerebral Cortex, 24*(6), 1619–1629. https://doi.org/10.1093/cercor/bht019

Hellerstedt, R., & Johansson, M. (2016). Competitive semantic memory retrieval: Temporal dynamics revealed by event-related potentials. *PLoS ONE, 11*(2), e0150091. https://doi.org/10.1371/journal.pone.0150091

Hermans, D. (2004). Between-language identity effects in picture-word interference tasks: A challenge for language-nonspecific or language-specific models of lexical access? *International Journal of Bilingualism, 8*(2), 115–125. https://doi.org/10.1177/13670069040080020101

Hermans, D., Bongaerts, T., De Bot, K., & Schreuder, R. (1998). Producing words in a foreign language: Can speakers prevent interference from their first language? *Bilingualism: Language and Cognition, 1*(3), 213–229. https://doi.org/10.1017/S1366728998000364

Higby, E., Lerman, A., Korytkowska, M., Malcolm, T., & Obler, L. (2019). Ageing as a confound in language attrition research. In M. S. Schmid & B. Köpke (Eds.), *The Oxford handbook of language attrition* (pp. 121–135). Oxford University Press.

Houston, J. P. (1967). Retroactive inhibition and point of interpolation. *Journal of Verbal Learning and Verbal Behavior, 6*(1), 84–88. https://doi.org/10.1016/S0022-5371(67)80054-9

Hulbert, J. C., & Norman, K. A. (2014). Neural differentiation tracks improved recall of competing memories following interleaved study and retrieval practice. *Cerebral Cortex, 25*(10), 3994–4008. https://doi.org/10.1093/cercor/bhu284

Hupbach, A., Gomez, R., Hardt, O., & Nadel, L. (2007). Reconsolidation of episodic memories: A subtle reminder triggers integration of new information. *Learning & Memory, 14*, 47–53. https://doi.org/10.1101/lm.365707

Ibrahim, A., Cowell, P. E., & Varley, R. A. (2017). Word frequency predicts translation asymmetry. *Journal of Memory and Language, 95*, 49–67. https://doi.org/10.1016/j.jml.2017.02.001

Isurin, L. (2000). Deserted island or a child's first language forgetting. *Bilingualism: Language and Cognition, 3*(2), 151–166. https://doi.org/10.1017/S1366728900000237

Isurin, L., & McDonald, J. L. (2001). Retroactive interference from translation equivalents: Implications for first language forgetting. *Memory & Cognition, 29*(2), 312–319. https://doi.org/10.3758/BF03194925

Jackson, G. M., Swainson, R., Cunnington, R., & Jackson, S. R. (2001). ERP correlates of executive control during repeated language switching. *Bilingualism: Language and Cognition, 4*(2), 169–178. https://doi.org/10.1017/S1366728901000268

Jakobson, R. (1941). *Kindersprache, Aphasie und allgemeine Lautgesetze*. Almqvist.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and phonological form. *Journal of Experimental Psychology: Learning Memory and Cognition, 20*(4), 824–843. http://dx.doi.org/10.1037/0278-7393.20.4.824

Johansson, M., Aslan, A., Baüml, K.-H., Gäbel, A., & Mecklinger, A. (2007). When remembering causes forgetting: Electrophysiological correlates of retrieval-induced forgetting. *Cerebral Cortex, 17*(6), 1335–1341. https://doi.org/10.1093/cercor/bhl044

Keppel, G., & Underwood, B. J. (1962). Proactive inhibition in short-term retention of single items. *Journal of Verbal Learning and Verbal Behavior, 1*(3), 153–161. https://doi.org/10.1016/S0022-5371(62)80023-1

Kindt, M., & Soeter, M. (2013). Reconsolidation in a human fear conditioning study: A test of extinction as updating mechanism. *Biological Psychology, 92*(1), 43–50. https://doi.org/10.1016/j.biopsycho.2011.09.016

Kleinman, D., & Gollan, T. H. (2018). Inhibition accumulates over time at multiple processing levels in bilingual language control. *Cognition, 173*, 115–132. https://doi.org/10.1016/j.cognition.2018.01.009

Klimesch, W., Doppelmayr, M., Russegger, H., & Pachinger, T. (1996). Theta band power in the human scalp EEG and the encoding of new information. *Neuroreport, 7*(7), 1235–1240. https://doi.org/10.1097/00001756-199605170-00002

Köpke, B. (2002). Activation thresholds and non-pathological first language attrition. In F. Fabbro (Ed.), *Advances in the neurolinguistics of bilingualism: Essays in honor of Michel Paradis* (pp. 119–142). Forum.

Köpke, B., & Keijzer, M. (2019). Introduction to psycholinguistic and neurolinguistic approaches to language attrition. In M. S. Schmid & B. Köpke (Eds.), *The Oxford handbook of language attrition* (pp. 63-72). Oxford University Press.

Köpke, B., & Schmid, M. S. (2004). Language attrition: The next phase. In M. S. Schmid, B. Köpcke, M. Keijzer, & L. Weilemar (Eds.), *First language attrition: Interdisciplinary perspectives on methodological issues* (pp. 1–43). John Benjamins.

Kroll, J. F., Bobb, S. C., Misra, M., & Guo, T. (2008). Language selection in bilingual speech: Evidence for inhibitory processes. *Acta Psychologica, 128*(3), 416–430. https://doi.org/10.1016/j.actpsy.2008.02.001

Kroll, J. F., Bobb, S. C., & Wodniecka, Z. (2006). Language selectivity is the exception, not the rule: Arguments against a fixed locus of language selection in bilingual speech. *Bilingualism: Language and Cognition, 9*(2), 119–135. https://doi.org/10.1017/S1366728906002483

Kroll, J. F., Bogulski, C. A., & McClain, R. (2012). Psycholinguistic perspectives on second language learning and bilingualism. *Linguistic Approaches to Bilingualism, 2*(1), 1–24. https://doi.org/10.1075/lab.2.1.01kro

Kroll, J. F., Gullifer, J. W., & Rossi, E. (2013). The multilingual lexicon: The cognitive and neural basis of lexical comprehension and production in two or more languages. *Annual Review of Applied Linguistics, 33*, 102–127. https://doi.org/10.1017/S0267190513000111

Kuhberg, H. (1992). Longitudinal L2-attrition versus L2-acquisition, in three Turkish children- empirical findings. *Interlanguage Studies Bulletin (Utrecht), 8*(2), 138–154. https://doi.org/10.1177/026765839200800203

Lambert, R. D., & Freed, B. (1982). *The loss of language skills.* Newbury House.

Landauer, T. K. (1974). Consolidation in human memory: Retrograde amnestic effects of confusable items in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior, 13*(1), 45–53. https://doi.org/10.1016/S0022-5371(74)80029-0

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

Lemhöfer, K., Schellenberger, J., & Schriefers, H. (2020). *Language conflict in trilingual word production: Evidence from the phono-translation effect*. [Manuscript submitted for publication]. Radboud University.

Lemhöfer, K., Spalek, K., & Schriefers, H. (2008). Cross-language effects of grammatical gender in bilingual word recognition and production. *Journal of Memory and Language, 59*(3), 312–330. https://doi.org/10.1016/j.jml.2008.06.005

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory, 10*(8), 707–710.

Levy, B. J., & Anderson, M. C. (2002). Inhibitory processes and the control of memory retrieval. *Trends in Cognitive Sciences, 6*(7), 299–305. https://doi.org/10.1016/S1364-6613(02)01923-X

Levy, B. J., McVeigh, N. D., Marful, A., & Anderson, M. C. (2007). Inhibiting your native language: The role of retrieval-induced forgetting during second language acquisition. *Psychological Science, 18*(1), 29–34. https://doi.org/10.1111/j.1467-9280.2007.01844.x

Linck, J. A., Hoshino, N., & Kroll, J. F. (2008). Cross-language lexical processes and inhibitory control. *The Mental Lexicon, 3*(3), 349–374. https://doi.org/10.1075/ml.3.3.06lin

Linck, J. A., & Kroll, J. F. (2019). Memory retrieval and language attrition: Language loss or manifestation of a dynamic system? In M. S. Schmid & B. Köpke (Eds.), *The Oxford handbook of language attrition* (pp. 88–97). Oxford University Press.

Lindsay, S., & Gaskell, M. G. (2013). Lexical integration of novel words without sleep. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(2), 608–622. https://doi.org/10.1037/a0029243

Llama, R., Cardoso, W., & Collins, L. (2010). The influence of language distance and language status on the acquisition of L3 phonology. *International Journal of Multilingualism, 7*(1), 39–57. https://doi.org/10.1080/14790710902972255

Luo, L., Luk, G., & Bialystok, E. (2010). Effect of language proficiency and executive control on verbal fluency performance in bilinguals. *Cognition, 114*(1), 29–41. https://doi.org/10.1016/j.cognition.2009.08.014

MacLeod, M. (2002). Retrieval-induced forgetting in eyewitness memory: Forgettting as a consequence of remembering. *Applied Cognitive Psychology, 16*(2), 135–149. https://doi.org/10.1002/acp.782

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language, 92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

McGaugh, J. L. (2000). Memory- a century of consolidation. *Science, 287*(5451), 248–251. https://doi.org/10.1126/science.287.5451.248

McGeoch, J. A., & Nolen, M. E. (1933). Studies in retroactive inhibition. IV. Temporal point of interpolation and degree of retroactive inhibition. *Journal of Comparative Psychology, 15*(3), 407–417. https://doi.org/10.1037/h0072751

McGeoch, J. A. (1932a). Forgetting and the law of disuse. *Psychological Review, 39*(4), 352–370. https://doi.org/10.1037/h0069819

McGeoch, J. A. (1932b). The influence of degree of interpolated learning upon retroactive inhibition. *The American Journal of Psychology, 44*(4), 695–708. https://doi.org/10.2307/1414532

Mehotcheva, T. H. (2010). *After the fiesta is over: Foreign language attrition of Spanish in Dutch and German Erasmus students* [Unpublished doctoral dissertation, Universitat Pompeu Fabra]. https://www.tdx.cat/bitstream/handle/10803/37468/ttm.pdf?sequence

Mehotcheva, T. H., & Köpke, B. (2019). Introduction to L2 attrition. In M. S. Schmid & B. Köpke (Eds.), *The Oxford handbook of language attrition* (pp. 331–348). Oxford University Press.

Mehotcheva, T. H., & Mytara, K. (2019). Exploring the impact of extralinguistic factors on L2/FL attrition. In M. S. Schmid & B. Köpke (Eds.), *The Oxford handbook of language attrition* (pp. 349–363). Oxford University Press.

Melinger, A. (2018). Distinguishing languages from dialects: A litmus test using the picture-word interference task. *Cognition, 172*, 73–88. https://doi.org/10.1016/j.cognition.2017.12.006

Meuter, R. F. I., & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of Memory and Language, 40*, 25–40. https://doi.org/10.1006/jmla.1998.2602

Mickan, A., McQueen, J. M., & Lemhöfer, K. (2019). Bridging the gap between second language acquisition research and memory science: The case of foreign language attrition. *Frontiers in Human Neuroscience, 13*, 397. https://doi.org/10.3389/fnhum.2019.00397

Misra, M., Guo, T., Bobb, S. C., & Kroll, J. F. (2012). When bilinguals choose a single word to speak: Electrophysiological evidence for inhibition of the native language. *Journal of Memory and Language, 67*(1), 224–237. https://doi.org/10.1016/j.jml.2012.05.001

Moorcroft, R., & Gardner, R. C. (1987). Linguistic factors in second-language loss. *Language Learning, 37*(3), 327–340. https://doi.org/10.1111/j.1467-1770.1987.tb00574.x

Müller, G. E., & Pilzecker, A. (1900). Experimentelle Beiträge zur Lehre vom Gedächtnis. *Zeitschrift Für Psychologie Ergänzungsband, 1*, 1–300.

Murtagh, L. (2003). *Retention and attrition of Irish as a second language* [Unpublished doctoral dissertation, University of Groningen]. https://www.rug.nl/research/portal/files/2999446/thesis.pdf

Nieuwenhuis, S., Yeung, N., van den Wildenberg, W., & Ridderinkhof, K. R. (2003). Electrophysiological correlates of anterior cingulate function in a go/no-go task: Effects of response conflict and trial type frequency. *Cognitive, Affective, & Behavioral Neuroscience, 3*(1), 17–26. https://doi.org/10.3758/CABN.3.1.17

Nigbur, R., Ivanova, G., & Stürmer, B. (2011). Theta power as a marker for cognitive interference. *Clinical Neurophysiology, 122*(11), 2185–2194. https://doi.org/10.1016/j.clinph.2011.03.030

Nikitina, L., & Furuoka, F. (2005). Integrative motivation in a foreign language classroom: A study on the nature of motivation of the Russian language learners in universiti Malaysia Sabah. *Jurnal Kinabalu, Jurnal Perniagaan & Sains Sosial, 11*, 23–34.

Olshtain, E. (1986). The attrition of English as a second language with speakers of Hebrew. In B. Weltens, K. de Bot, & T.J.M. van Els (Eds.), *Language attrition in progress* (pp. 187–204). Foris.

Olshtain, E. (1989). Is second language attrition the reversal of second language acquisition? *Studies in Second Language Acquisition, 11*(2), 151–165. https://doi.org/10.1017/S0272263100000589

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience, 2011*, 156869. https://doi.org/10.1155/2011/156869

Osterhout, L., Pitkänen, I., McLaughlin, J., & Zeitlin, M. (2019). Event-related potentials as metrics of foreign language learning and loss. In M. S. Schmid & B. Köpke (Eds.), *The Oxford handbook of language attrition* (pp. 403–416). Oxford University Press.

Pallier, C., Dehaene, S., Poline, J.-B., LeBihan, D., Argenti, A.-M., Dupoux, E., & Mehler, J. (2003). Brain imaging of language plasticity: Can a second language replace the first? *Cerebral Cortex, 13*(2), 155–161. https://doi.org/10.1093/cercor/13.2.155

Pan, B. A., & Berko-Gleason, J. (1986). The study of language loss: Models and hypotheses for an emerging discipline. *Applied Psycholinguistics, 7*(3), 193-206. https://doi.org/10.1017/S0142716400007530

Paradis, M. (1993). Linguistic, psycholinguistic, and neurolinguistic aspects of “interference” in bilingual speakers: The activation threshold hypothesis. *International Journal of Psycholinguistics, 9*(2), 133–145.

Paradis, M. (2004). *A neurolinguistic theory of bilingualism.* John Benjamins. https://doi.org/10.1075/sibil.18

Paradis, M. (2007). L1 attrition features predicted by a neurolinguistic theory of bilingualism. In B. Köpke, M. S. Schmid, M. Keijzer, & S. Bezdjian (Eds.), *Language attrition: Theoretical perspectives* (pp. 121–133). John Benjamins.

Penolazzi, B., Stramaccia, D. F., Braga, M., Mondini, S., & Galfano, G. (2014). Human memory retrieval and inhibitory control in the brain: Beyond correlational evidence. *Journal of Neuroscience, 34*(19), 6606–6610. https://doi.org/10.1523/JNEUROSCI.0349-14.2014

Piai, V., Roelofs, A., Jensen, O., Schoffelen, J.-M., & Bonnefond, M. (2014). Distinct patterns of brain activity characterise lexical activation and competition in spoken word production. *PLoS ONE, 9*(2), e88674. https://doi.org/10.1371/journal.pone.0088674

Pierce, L. J., Klein, D., Chen, J.-K., Delcenserie, A., & Genesee, F. (2014). Mapping the unconscious maintenance of a lost first language. *Proceedings of the National Academy of Sciences, 112*(8), 17314–17319. https://doi.org/10.1073/pnas.1409411111

Pöhlchen, D., Pawlizki, A., Gais, S., & Schönauer, M. (2020). Evidence against a large effect of sleep in protecting verbal memories from interference. *Journal of Sleep Research*, e13042. https://doi.org/10.1111/jsr.13042

Poort, E. D., Warren, J. E., & Rodd, J. M. (2016). Recent experience with cognates and interlingual homographs in one language affects subsequent processing in another language. *Bilingualism, 19*(1), 206–212. https://doi.org/10.1017/S1366728915000395

Postman, L., & Alper, T. G. (1946). Retroactive inhibition as a function of the time of interpolation of the inhibitor between learning and recall. *The American Journal of Psychology, 59*(3), 439–449. https://doi.org/10.2307/1417613

Postman, L., & Kaplan, H. L. (1947). Reaction time as a measure of retroactive inhibition. *Journal of Experimental Psychology, 37*(2), 136–145. https://doi.org/10.1037/h0055703

Potts, R., & Shanks, D. R. (2012). Can testing immunize memories against interference? *Journal of Experimental Psychology. Learning, Memory, and Cognition, 38*(6), 1780–1785. https://doi.org/10.1037/a0028218

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Raaijmakers, J. G. W., & Jakab, E. (2013). Rethinking inhibition theory: On the problematic status of the inhibition theory for forgetting. *Journal of Memory and Language, 68*(2), 98–122. https://doi.org/10.1016/j.jml.2012.10.002

Reetz-Kurashige, A. (1999). Japanese returnees' retention of English-speaking skills: Changes in verb usage over time. In L. Hansen (Ed.), *Second language attrition in Japanese contexts* (pp. 21–58). Oxford University Press.

Reppa, I., Williams, K. E., Worth, E. R., Greville, W. J., & Saunders, J. (2017). Memorable objects are more susceptible to forgetting: Evidence for the inhibitory account of retrieval-induced forgetting. *Acta Psychologica, 181*, 51–61. https://doi.org/10.1016/j.actpsy.2017.09.012

Roediger, H. L. (1973). Inhibition in recall from cueing with recall targets. *Journal of Verbal Learning and Verbal Behavior, 12*(6), 644–657. https://doi.org/10.1016/S0022-5371(73)80044-1

Roediger, H. L., Weinstein, Y., & Agarwal, P. (2010). Forgetting: Preliminary considerations. In S. Della Sala (Ed.), *Forgetting* (pp. 15–36). Psychology Press.

Román, P., Soriano, M. F., Gómez-Ariza, C. J., & Bajo, M. T. (2009). Retrieval-induced forgetting and executive control. *Psychological Science, 20*(9), 1053–1058. https://doi.org/10.1111/j.1467-9280.2009.02415.x

Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high-and low-frequency words. *Memory & Cognition, 18*(4), 367–379. https://doi.org/10.3758/BF03197126

Rugg, M. D., Cox, C. J. C., Doyle, M. C., & Wells, T. (1995). Event-related potentials and the recollection of low and high frequency words. *Neuropsychologia, 33*(4), 471–484. https://doi.org/10.1016/0028-3932(94)00132-9

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Sciences, 11*(6), 251–257. https://doi.org/10.1016/j.tics.2007.04.004

Runnqvist, E., & Costa, A. (2012). Is retrieval-induced forgetting behind the bilingual disadvantage in word production? *Bilingualism: Language and Cognition, 15*(2), 365–377. https://doi.org/10.1017/S1366728911000034

Ruts, W., de Deyne, S., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behaviour Research Methods, Instruments, & Computers, 36*(3), 506-515.

Sanders, A. F., Whitaker, L., & Cofer, C. N. (1974). Evidence for retroactive interference in recognition from reaction time. *Journal of Experimental Psychology, 102*(6), 1126–1129. https://doi.org/10.1037/h0036380

Schiller, D., Monfils, M.-H., Raio, C. M., Johnson, D. C., LeDoux, J. E., & Phelps, E. A. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature, 463*(7277), 49–53. https://doi.org/10.1038/nature08637

Schmid, M. S. (2007). The role of L1 use for L1 attrition. In B. Köpke, M. S. Schmid, M. Keijzer, & S. Dostert (Eds.), *Language attrition: Theoretical perspectives* (pp. 135–153). John Benjamins.

Schmid, M. S. (2019). The impact of frequency of use and length of residence on L1 attrition. In M. S. Schmid & B. Köpke (Eds.), *The Oxford handbook of language attrition* (pp. 288–303). Oxford University Press.

Schmid, M. S. (2016). First language attrition. *Language Teaching, 49*(2), 186–212. https://doi.org/10.1017/S0261444815000476

Schmid, M. S., & Keijzer, M. (2009). First language attrition and reversion among older migrants. *International Journal of the Sociology of Language, 2009*(200), 83–101. https://doi.org/10.1515/IJSL.2009.046

Schmid, M. S, & Köpke, B. (2019). *The Oxford handbook of language attrition*. Oxford University Press.

Schmid, M. S., & Mehotcheva, T. H. (2012). Foreign language attrition. *Dutch Journal of Applied Linguistics, 1*(1), 102–124. https://doi.org/10.1075/dujal.1.1.08sch

Seliger, H. W., & Vago, R. M. (1991). The study of first language attrition: An overview. In H. W. Seliger & R. M. Vago (Eds.), *First language attrition* (pp. 3–15). Cambridge University Press.

Shao, Z., Janse, E., Visser, K., & Meyer, A. S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology, 5*, 772. https://doi.org/10.3389/fpsyg.2014.00772

Sharwood Smith, M. (1989). Crosslinguistic influence in language loss. In *Bilingualism across the lifespan: Aspects of acquisition, maturity, and loss* (pp. 185 – 201). Cambridge University Press.

Sheth, B. R., Varghese, R., & Thuy Truong, B. S. (2012). Sleep shelters verbal memory from different kinds of interference. *Sleep, 35*(7), 985–996. https://doi.org/10.5665/sleep.1966

Sisson, E. D. (1939). Retroactive inhibition: The temporal position of interpolated activity. *Journal of Experimental Psychology, 25*(2), 228–233. https://doi.org/10.1037/h0055178

Skaggs, E. B. (1925). Further studies in retroactive inhibition. *Psychology Monograph, 34*(8), 1–60.

Smith, M. E. (1993). Neurophysiological manifestations of recollective experience during recognition memory judgments. *Journal of Cognitive Neuroscience, 5*(1), 1–13. https://doi.org/10.1162/jocn.1993.5.1.1

Staudigl, T., Hanslmayr, S., & Bäuml, K.-H. T. (2010). Theta oscillations reflect the dynamics of interference in episodic memory retrieval. *Journal of Neuroscience, 30*(34), 11356–11362. https://doi.org/10.1523/JNEUROSCI.0637-10.2010

Stickgold, R., & Walker, M. P. (2005). Memory consolidation and reconsolidation: What is the role of sleep? *Trends in Neurosciences, 28*(8), 408–415. https://doi.org/10.1016/j.tins.2005.06.004

Storm, B. C., Angello, G., Buchli, D. R., Koppel, R. H., Little, J. L., & Nestojko, J. F. (2015). A review of retrieval-induced forgetting in the contexts of learning, eyewitness memory, social cognition, autobiographical memory, and creative cognition. In B. H. Moss (Ed.), *The psychology of learning and motivation* (Vol. 62, pp. 141–194). Elsevier. https://doi.org/10.1016/bs.plm.2014.09.005

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2012). On the durability of retrieval-induced forgetting. *Journal of Cognitive Psychology, 24*(5), 617–630. https://doi.org/10.1080/20445911.2012.674030

Storm, B. C., Bjork, E. L., Bjork, R. A., & Nestojko, J. F. (2006). Is retrieval success a necessary condition for retrieval-induced forgetting? *Psychonomic Bulletin and Review, 13*(6), 1023–1027. https://doi.org/10.3758/BF03213919

Strange, B. A., Kroes, M. C., Fan, J., & Dolan, R. J. (2010). Emotion causes targeted forgetting of established memories. *Frontiers in Behavioral Neuroscience, 4*, 175. https://doi.org/10.3389/fnbeh.2010.00175

Takashima, A., Petersson, K. M., Rutters, F., Tendolkar, I., Jensen, O., Zwarts, M. J., & McNaughton, B. L. (2006). Declarative memory consolidation in humans: A prospective functional magnetic resonance imaging study. *Proceedings of the National Academy of Sciences of the United States of America, 103*(3), 756–761. https://doi.org/10.1073/pnas.0507774103

Taura, H. (2008). *Language attrition and retention in Japenese returnee students.* Akashi Shoten.

Tempel, T., & Frings, C. (2013). Resolving interference between body movements: Retrieval-induced forgetting of motor sequences. *Journal of Experimental Psychology: Learning Memory and Cognition, 39*(4), 1152–1161. https://doi.org/10.1037/a0030336

Thorndike, E. L. (1914). *The psychology of learning.* Teacher College.

Tomiyama, M. (2000). Child second language attrition: A longitudinal case study. *Applied Linguistics, 21*(3), 304–332. https://doi.org/10.1093/applin/21.3.304

Tomiyama, M. (2008). Age and proficiency in L2 attrition: Data from two siblings. *Applied Linguistics, 30*(2), 253–275. https://doi.org/10.1093/applin/amn038

Treccani, B., Argyri, E., Sorace, A., & Della Sala, S. (2009). Spatial negative priming in bilingualism. *Psychonomic Bulletin & Review, 16*(2), 320–327. https://doi.org/10.3758/PBR.16.2.320

Van Hell, J. G., & Tanner, D. (2012). Second language proficiency and cross-language lexical activation. *Language Learning, 62*(s2), 148–171. https://doi.org/10.1111/j.1467-9922.2012.00710.x van Kesteren, M. T. R., Rijpkema, M., Ruiter, D. J., & Fernandez, G. (2010). Retrieval of associative information congruent with prior knowledge is related to increased medial prefrontal activity and connectivity. *Journal of Neuroscience, 30*(47), 15888–15894. https://doi.org/10.1523/JNEUROSCI.2674-10.2010

van Overschede, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language, 50*(3), 289-335. https://doi.org/10.1016/j.jml.2003.10.003

Walker, M. P., Brakefield, T., Hobson, J. A., & Stickgold, R. (2003). Dissociable stages of human memory consolidation and reconsolidation. *Nature, 425*, 616–620. https://doi.org/10.1038/nature01930

Wang, X. (2010). *Patterns and causes of attrition of English as a foreign language.* [Unpublished doctoral dissertation, Shandong University]. http://www.let.rug.nl/languageattrition/Papers/Wang2010.pdf

Weltens, B. (1988). *The attrition of French as a foreign language* [Unpublished doctoral dissertation, Katholieke Universiteit Nijmegen]. https://hdl.handle.net/2066/113589

Weltens, B., Van Els, T. J. M., & Schils, E. (1989). The long-term retention of French by Dutch students. *Studies in Second Language Acquisition, 11*(2), 205–216. https://doi.org/10.1017/S0272263100000619

Wichert, S., Wolf, O. T., & Schwabe, L. (2011). Reactivation, interference, and reconsolidation: Are recent and remote memories likewise susceptible? *Behavioral Neuroscience, 125*(5), 699–704. https://doi.org/10.1037/a0025235

Wilding, E. L. (2000). In what way does the parietal ERP old/new effect index recollection? *International Journal of Psychophysiology, 35*(1), 81–87. https://doi.org/10.1016/S0167-8760(99)00095-1

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

Williams, S., & Hammarberg, B. (1998). Language switches in L3 production: Implications for a polyglot speaking model. *Applied Linguistics, 19*(3), 295–333. https://doi.org/10.1093/applin/19.3.295

Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature Neuroscience, 18*(4), 582–589. https://doi.org/10.1038/nn.3973

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology, 55*, 235–269. https://doi.org/10.1146/annurev.psych.55.090902.141555

Wodniecka, Z., Szewczyk, J., Kałamała, P., Mandera, P., & Durlik, J. (2020). When a second language hits a native language. What ERPs (do and do not) tell us about language retrieval difficulty in bilingual language production. *Neuropsychologia, 141*, 107390. https://doi.org/10.1016/j.neuropsychologia.2020.107390

Xu, X. (2010). *English language attrition and retention in Chinese and Dutch university students* [Unpublished doctoral dissertation]. University of Groningen.

Yoshitomi, A. (1999). On the loss of English as a L2 by Japenese returnee children. In L. Hansen (Ed.), *Second language attrition in Japense contexts* (pp. 80–111). Oxford University Press.

Zhang, Q., Popov, V., Koch, G. E., Calloway, R. C., & Coutanche, M. N. (2018). Fast memory integration facilitated by schema consistency. *BioRxiv*, 253393. https://doi.org/10.1101/253393

Zheng, X., Roelofs, A., Erkan, H., & Lemhöfer, K. (2020). Dynamics of inhibitory control during bilingual speech production: An electrophysiological study. *Neuropsychologia, 140*, 107387. https://doi.org/10.1016/j.neuropsychologia.2020.107387

Appendices

APPENDIX A

Supplementary materials for Chapter 2

A.1 | Stimulus Database for Experiment 1

English	Dutch	Spanish
iron	strijkijzer	plancha
carrot	wortel	zanahoria
bush	struik	arbusto
needle	naald	aguja
swing	schommel	columpio
dice	dobbelsteen	dado
mushroom	paddenstoel	hongo
root	wortel	raíz
doll	pop	muñeca
keyboard	toetsenbord	teclado
earring	oorbel	pendiente
bracelet	armband	pulsera
scarf	sjaal	bufanda
sleeve	mouw	manga
zipper	rits	cremallera
chain	ketting	cadena
whistle	fluit	silbato
steeringwheel	stuur	volante
raincoat	regenjas	gabardina
moustache	snor	bigote
broom	bezem	escoba
pillow	kussen	almohada
mop	dweil	fregona
peach	perzik	melocotón
squirrel	eekhoorn	ardilla
donkey	ezel	burro
ant	mier	hormiga
bowl	kom	cuenco
ashtray	asbak	cenicero
bullet	kogel	bala
arrow	pijl	flecha
lighter	aansteker	mechero
bone	bot	hueso
wave	golf	ola
shark	haai	tiburón
cage	kooi	jaula
fence	hek	alambrada
brick	baksteen	ladrillo

English	Dutch	Spanish
straw	rietje	paja
butterfly	vlinder	mariposa
egg	ei	huevo
mirror	spiegel	espejo
key	sleutel	llave
dress	jurk	vestido
painting	schilderij	cuadro
cow	koe	vaca
spoon	lepel	cuchara
strawberry	aardbei	fresa
toothbrush	tandenborstel	cepillo de dientes
knife	mes	cuchillo
glasses	bril	gafas
watch	horloge	reloj
candle	kaars	vela
onion	ui	cebolla
spider	spin	araña
peanut	pinda	cacahuete
duck	eend	pato
wood	hout	madera
skirt	rok	falda
goat	geit	cabra
sausage	worst	salchicha
refrigerator	koelkast	frigorífico
lipstick	lippenstift	pintalabios
bridge	brug	puente
wallet	portemonnee	cartera
suit	pak	traje
monkey	aap	mono
garlic	knoflook	ajo
stamp	postzegel	sello
bra	beha	sujetador
microwave	magnetron	microondas
cloud	wolk	nube
pencil	potlood	lápiz
orange	sinaasappel	naranja
umbrella	paraplu	paraguas
trafficlight	stoplicht	semáforo
rope	touw	soga
frog	kikker	rana
tie	stropdas	corbata
bucket	emmer	cubo
shell	schelp	concha
backpack	rugzak	mochila

English	Dutch	Spanish
snail	slak	caracol
suitcase	koffer	maleta
hairbrush	haarborstel	cepillo
fingernail	vingernagel	uña
coin	munt	moneda
nail	nagel	clavo
fly	vlieg	mosca
scissors	schaar	tijeras
grapes	druiven	uvas
pigeon	duif	paloma
lock	slot	cerradura
bow	boog	arco
belt	riem	cinturón
glove	handschoen	guante
parrot	papegaai	loro
shovel	schep	pala
blanket	deken	manta
axe	bijl	hacha
mower	grasmaaier	cortacésped
jar	pot	tarro
paintbrush	kwast	pincel
basket	mand	canasta
saw	zaag	sierra
coat	jas	abrigo
headphones	koptelefoon	cascos
match	lucifer	cerillo
razor	scheermes	cuchilla
slide	glijbaan	tobogán
screwdriver	schroevendraaier	destornillador
screw	schroef	navaja
hairdryer	föhn	secador de pelo
branch	tak	rama
eggplant	aubergine	berenjena
leash	hondenlijn	correa del perro
diaper	luier	pañal
kite	vlieger	cometa
vacuumcleaner	stofzuiger	aspiradora
lid	deksel	tapa
hat	muts	gorro
drill	boor	taladro
stapler	nietmachine	grapadora
pencilsharpener	puntenslijper	sacapuntas
seagull	meeuw	gaviota
cake	taart	pastel

English	Dutch	Spanish
chest	kist	baúl
ruler	liniaal	regla
sled	slee	trineo
cuttingboard	snijplank	tabla de cortar
scale	weegschaal	balanza
bowtie	strik	pajarita
tap	kraan	grifo
chalk	krijt	tiza
stool	kruk	taburete
alarmclock	wekker	despertador
beetle	kever	escarabajo
cradle	wieg	cuna
cane	stok	bastón
antlers	gewei	cornamenta
frame	lijst	marco
purse	tas	bolso
socket	stopcontact	enchufe
rake	hark	rastrillo
radiator	verwarming	calefacción
marble	knikker	canica
plunger	ontstopper	desatascador
rooster	haan	gallo
fan	waaier	abanico
apron	schort	delantal
cap	pet	gorra
clog	klomp	zueco
coaster	onderzetter	posavasos
acorn	eikel	bellota
pacifier	fopspeen	chupete
buckle	gesp	hebilla
scooter	step	patinete
bellpepper	paprika	pimiento
flyswatter	vliegenmepper	matamoscas
wardrobe	kast	armario
hinge	scharnier	bisagra
seesaw	wip	balancín
bib	slab	babero
wheelbarrow	kruiwagen	carretilla
yarn	garen	hilo
wateringcan	gieter	regadera
top	tol	peonza
peg	wasknijper	pinza
shoehorn	schoenlepel	calzador

Note. Items are in the order as presented in the pre-test.

A.2 | Item Selection Details

Each participant's final set of 40 experimental items consisted of two subsets: 20 words that would receive interference on day 2 and 20 that would not. Experimental items for these sets were chosen based on the participant's performance in the pre-test. As explained in the main manuscript, the initial 40 words in the pre-test served as the ideal base set. If a participant already knew words from this set, these words had to be replaced with unknown words from the remaining pre-test items. Items in the base set were pre-assigned to either one of the two interference subsets, and replacements inherited the subset-assignments from the words they replaced. Which of the two subsets ultimately received interference was counterbalanced across participants. Importantly, words in these two subsets were matched on a number of criteria, each of which will be briefly explained and motivated below.

A.2.1 | *Matching Criteria*

A.2.1.1 | Experimental Items

Spanish word length was measured in syllables, and was controlled for to ensure that words in both subsets were equally difficult to learn.

Phonological similarity was assessed via Levensthein distances: words within the final set had to be at least two Levenshtein distance units (Levenshtein, 1966) apart from one another, so as to avoid confusion during learning and subsequent recall.

Semantic similarity was controlled for by means of semantic vectors (CBOW space, English lemmas) and their cosine distances as reported and explained in Mandera et al. (2017). The cosine distances were obtained from their open source web interface (http://meshugga.ugent.be/snaut/). Distances between the semantic vectors for any two given words indicate how semantically similar they are, with small values indicating small distances, and thus high similarity (zero would be the value for two identical vectors). We restricted semantic similarity across subsets such that each word had to be a minimal semantic distance apart from all words in the other subset. In other words, items with a low semantic distance, such as 'key' and 'lock', would be fine within the same subset, but would not be allowed in different subsets for a given participant. Interference resulting from the retrieval of one of these two items may spread to the other, given their semantic association, which would potentially weaken or wash out the language interference effect if these two items were in different subsets.

For a similar reason, we also controlled for semantic similarity *within* subsets. We avoided, as much as possible, that one subset consisted of only members of one semantic category (resulting in a low average within set semantic distance, and thus likely additional semantic interference among items), while the other subset had items from many different categories (resulting in a high average distance, and less additional semantic interference among items). Average within-subset similarity was not allowed to differ statistically between subsets. This constraint was introduced to avoid differences in learning difficulty, as well as to preclude differences between the subsets in the amount of interference among items (i.e., avoiding the above scenario of having more additional, semantic interference in one compared to the other subset).

A.2.1.2 | Filler Items

Filler items (for the interference tasks on day 2) were chosen from the remaining 129 items (i.e., those that were not selected as target items). Similar semantic relatedness constraints as for the target items were imposed: a filler needed to be a certain minimal semantic distance away from all target items (0.7), and its average semantic distance to items of each set had to be roughly the same. Spanish word length and Levensthein distance did not need to be controlled for, since fillers were only named in Dutch or English.

A.2.2 | Script Logic

The Matlab (v.8.6, R2015b, The Math Works, Inc.) script replaced words one by one. For each to-be-replaced (already known in Spanish) word, the script initially chose the word that was semantically closest to it, that is, the replacement option with the smallest semantic distance to the to-be-replaced word (in case of multiple options, the script randomly picked one). Subsequently, the script checked whether this replacement option was within +-1 syllable length in Spanish from the to-be-replaced word, whether its Levenshtein distance with other words in the entire set was at least 2, and whether its semantic distance to words in the other subset would not exceed a predefined threshold of 0.68. Moreover, when the to-be-replaced word was a compound in either Dutch or English, the script would first attempt to find a replacement that was also a compound in either of those languages. Spanish compounds were not considered as experimental items, only as filler items. When a replacement option did not meet one or more of those criteria, it was removed from the list of possible replacements and the procedure was repeated until a viable replacement candidate was found. This way of replacing was chosen to ensure maximal overlap in item sets between participants.

On top of those restrictions, the two subsets of items had to be matched on a number of features. Once a replacement had been found it was added to the respective subset and the following aspects were evaluated:

- Word length in Spanish (in syllables): the average word length for the two subsets had to be roughly the same, i.e., not statistically different.
- Semantic similarity within subsets: the mean semantic similarity between items within each set could not be statistically different.

For all participants the script initially searched for replacements within the first 101 replacement options, as those were likely to be known in English (based on their frequency), only if that failed did I allow the script to search the full set of 169 items.

For eight out of the 54 tested participants, the script failed to find replacements for some of the already known words from the base set. In those cases, a more lenient script was used for item selection, sacrificing first the initial across-set semantic similarity constraint (successful for three participants), and second sacrificing the within-set semantic similarity constraint (successful for one additional participant). For four participants even the most lenient script did not succeed in compiling a final item set and they had to be sent home. Those were participants that already knew too many words in Spanish, thus resulting in a very limited set of replacement options.

A.3 | Filler Item Characteristics in Experiments 1 and 2

Table A.3.1
Filler item characteristics in Experiments 1 and 2.

	Experiment 1						Experiment 2					
	English			Dutch			Low frequency			High frequency		
	M	SD	range	M	SD	range	M	SD	range	M	SD	range
Dutch word length (in syllables)	1.79	0.88	1-4	1.72	0.87	1-4	1.85	0.93	1-4	1.20	0.41	1-2
English word length (in syllables)	1.69	0.73	1-3	1.71	0.76	1-3	2.1	1.12	1-6	1.40	0.60	1-3
Dutch Celex log frequency	1.02	0.67	0-2.14	1.03	0.67	0-2.14	0.34	0.36	0-0.85	1.65	0.50	1.04-2.91
Dutch Celex per million frequency	26.30	33.26	0-158	25.77	30.86	0-137	2.80	2.69	0-9	100.15	188.26	9-820

Note. In Experiment 1 item sets differed across participants. Means (*M*) and standard deviations (*SD*) were first calculated per subject and subsequently averaged over groups. Ranges show the absolute min and max

A.4 | Procedural Details for Each Task

A.4.1 | *Pre-test*

A trial started with a 500 ms blank screen, followed by the presentation of the picture in the center of the screen. Participants were given all the time they needed to provide their answer. The experimenter coded the answers for correctness by pressing one of two keys: ENTER for incorrect and SPACE for correct. As soon as the experimenter had coded the answer, the correct Spanish word appeared on screen underneath the picture (Verdana 30 pt, black, centered at 250 px below the center of the screen). The participant was then prompted to indicate whether they recognized the word or not, in which case the experimenter coded the word as correct (SPACE). If they had previously correctly named the word, the experimenter immediately coded it as correct, and when they said they did not know the word, the experimenter pressed ENTER. It was this last key press which was registered in the logsheet which was then passed to the Matlab script for item selection (see section A.2).

A.4.2 | *Learning Phase Tasks*

A.4.2.1 | **Familiarization - Self-Paced Learning**

Each trial started with a 500 ms blank screen, followed by the presentation of a picture (centered at 100 px above the center of the screen). After a 300 ms delay, the corresponding Spanish word appeared on screen (Verdana 30 pt, black, centered at 200 px below the center of the screen). After another 300 ms delay, the audio recording was played. The participant was then instructed to repeat the word out loud, and to initiate the next trial by clicking on the right arrow key whenever ready.

A.4.2.2 | **Two Alternative Forced Choice Task**

Each trial started with a 500 ms blank screen, followed by the simultaneous presentation of a picture (centered at 100 px above the center of the screen) and two Spanish words (Verdana 25 pt, black, horizontally at 200 px to the left and right of the center, and vertically at 200 px below the center). Both words were surrounded by a black circle each with a diameter of 216 px and a line width of 13 px. The participant subsequently had as much time as needed to choose the word that belongs to the picture. The choice was made via a mouse click within the circle of the corresponding word. A participant's response was followed by a 500 ms blank screen, which in turn was followed by a feedback screen. For feedback the circle surrounding the word that was clicked turned either green (when the answer was correct) or red. This color feedback was displayed for 500 ms, after which only the picture and the correct word were displayed together in the center of the screen and the corresponding audio was played. The picture and its label stayed on screen for another 500 ms after the end of the audio recording, after which the next trial started automatically.

In the second round, everything was identical with the only difference being that participants were first asked to attempt to recall the correct Spanish word. The experimenter coded their answers for correctness by pressing one of three keys (SPACE for correct, ENTER for incorrect, N for partially correct). Participants were allowed to take their time, but were also told that it was perfectly fine if they did not know the word yet. Immediately after the experimenter's key press, the two word options appeared on screen and the trial continued as in the first round.

A.4.2.3 | **Word Completion**

A trial started with a 500 ms blank screen, followed by the presentation of a picture (centered at 100 px above the center of the screen) and the first letter (or grapheme) of the corresponding Spanish word (Verdana 30 pt, black, centered at 200 px below the center of the screen). Participants were given no time limit to complete the word. Once the participant had made a naming attempt, the experimenter coded the correctness of the answer by pressing one of three buttons (same as for 2AFC). Based on the experimenter's coding, feedback was initiated immediately after the experimenter's key press: either a red or a green screen appeared around the picture, together with the full Spanish label and the spoken word. After a delay of 100 ms after the end of the audio recording, participants could move on to the next trial by clicking on the right arrow key. They were thus allowed to spend as much time as they needed on the feedback screen.

A.4.2.4 | **Writing**

A trial started with a 500 ms blank screen, followed by the presentation of the picture (centered at 100 px above the center of the screen). Participants then got as much time as needed to write the Spanish word down on a piece of paper. Once they had done so, they were instructed to click on the right arrow key in order to see and hear the correct Spanish word on screen (Verdana 24 pt, black, centered at 200 px below the center of the screen). Another click on the right arrow would start the next trial, but only after the participant had corrected him/herself with a red pen. Again, the next trial was self-initiated, providing as much time with the feedback screen as the participant needed.

A.4.2.5 | **Adaptive Picture Naming**

Again, each trial started with a 500 ms blank screen, followed by the presentation of the picture (centered at 100 px above the center of the screen). Participants then got as much time as they needed to name the picture in Spanish. The experimenter coded their answers, again via a key press, which started the feedback (again a green or red frame around the picture, the label (Verdana 24 pt, black, centered at 200 px below the center of the screen) underneath and the spoken word presented simultaneously). The participant could then initiate the next trial by clicking the right arrow key when ready, but at the earliest 100 ms after the end of the spoken word.

A.4.2.6 | Final Recall Tests

In all final recall tests, a trial started with a 500 ms blank screen, followed by the presentation of a picture (in the center of the screen). For the first Spanish post-test, immediately after learning on day 1, participants had a maximum of 30 seconds to respond. For the two final recall tests after interference (on day 2 and 8), participants did *not* have a time limit to give their answers. The experimenter coded their answers for correctness (same coding scheme as in 2AFC), and in doing so initiated the next trial (no further delay), but no feedback was provided to the participants.

A.4.3 | *Interference Phase Tasks*

A.4.3.1 | Familiarization

Each trial started with a 500 ms blank screen, followed by the presentation of the picture (centered at 100 px above the center of the screen) and the first letters of the English or Dutch word (Verdana 30 pt, black, centered at 200 px below the center of the screen). Participants then had as much time as needed to say the English/Dutch word out loud. The experimenter coded their answers for correctness (same coding as for learning tasks), which initiated the presentation of the full English/Dutch word on screen. People in the English group could then indicate whether they recognized the word or not, in case they had not named it properly initially. The experimenter again coded these answers, which then started the next trial.

A.4.3.2 | Picture Naming

Each trial started with a 500 ms blank screen, followed by the presentation of the picture in the middle of the screen. Participants were given unlimited time to name the picture in English/Dutch and the experimenter coded their answers for correctness (same coding as for all earlier tasks). Immediately after the experimenter’s key press the next trial started.

A.4.3.3 | Letter Search

Each trial started with a 500 ms blank screen, followed by the presentation of the picture in the middle of the screen. Participants then had 10 s to press a button corresponding to whether they thought the English/Dutch label for the picture contained a certain letter or not. Answers were given via a button box with two labeled buttons (‘Yes’, ‘No’; the ‘Yes’ button was always on the dominant hand). Immediately after the button press the next trial started.

A.4.4 | *Filler Tasks*

A.4.4.1 | **Simon Task**

In the Simon task participants had to respond to the color of a rectangle on the screen. They had to press either the right or the left arrow key depending on the color of the rectangle (color – button assignment was counterbalanced across participants). To make this task harder, the rectangle could either occur on the left or right side of the screen, and thus be either congruent (facilitating) or incongruent (distracting) with the position of the answer button.

A trial started with a 500 ms fixation cross (Arial, 20 pt, black) in the center of the screen, followed by a colored (red or blue) rectangle on either the left or right side of the screen. The rectangle remained visible for 500 ms. Participants had 1000 ms to press a button (500 ms during pic presentation + 500 ms with a white screen). After the button press there was a random delay of 0 to 500 ms, before the next trial started. The task started with 20 practice trials, during which participants received feedback (on screen for 500 ms, ‘Correct’, ‘Incorrect’, in either green or red, Arial 20 pt), and were told if they were too slow. After the practice round there were a total of 100 test trials (50 congruent (25 red, 25 blue), 50 incongruent (25 red, 25 blue), in randomized order). There were no breaks in the task.

The Simon effect is calculated as the difference in reaction times between the congruent and incongruent trials (considering correct trials only). The task was taken from the Experiment factory Github repository (https://github.com/expfactory-experiments/simon), which offers a set of cognitive tasks programmed in JavaScript.

A.4.4.2 | **Go-Nogo Task**

In the Go-NoGo task, participants had to press the space bar when they saw a rectangle with a specific color (orange or blue, counterbalanced across participants), while having to withhold the button press when the rectangle was in the other color (orange or blue respectively). Each trial started with a fixation cross (Arial, 20 pt, black) in the center of the screen for 500 ms, followed by a colored rectangle for 750 ms. Within those 750 ms participants had to respond by either pressing the space bar or withholding the button press. After a delay of 50 ms after either the maximum time to respond or the participant’s button press, the next trial started. The task started with 10 practice trials, during which the participant received feedback (shown on screen for 500 ms, ‘Correct’ or ‘Incorrect’, in green or red, Arial 20 pt) and was told if they were too slow (i.e., did not respond within 750 ms for the go trials). After each feedback screen, a blank screen was shown for 250 ms before the next trial started. After the practice round, there were a total 350 test trials (35 no-go, 315 go, in randomized order). There were no breaks in the task.

The NoGo effect is typically measured as the false alarm rate, i.e., the percentage of button presses when not called for. Again, the task was taken from the Experiment factory online repository (https://github.com/expfactory-experiments/go-nogo). The order of the two filler tasks was counterbalanced across participants.

A.5 | Accuracy Scoring

Participants’ Spanish word productions were compared to target (i.e., Spanish native speakers’) productions based on phonological similarity. Because we did not train our participants on pronunciation and because the focus of the present study is on lexical knowledge rather than pronunciation ability, common Dutch-Spanish orthography-based mispronunciation errors were ignored and counted as correct. These included the consistent mispronunciation of /u/ as /Y/, /η/ (written ‘ñ’ in Spanish) as /n/, /x/ (written ‘j’ in Spanish) as /j/, /ʎ/ (written ‘ll’ in Spanish) as /l/, and the pronunciation of ‘h’ while it should be silent. Likewise, a participant’s failure to pronounce intervocalic ‘c, ‘d’, ‘b’ and ‘g’ as /θ/, /ð/, /β/ and /γ/ respectively was ignored. The same goes for the realization of a trilled /r/, a phoneme that is inherently difficult to pronounce for many native speakers of Dutch. Finally, when participants needed multiple attempts to name a picture, the last production was used for scoring.

With these constraints in mind, and as explained in the main text in Chapter 2, responses were scored at the phoneme level. Incorrect phonemes could be either omissions, insertions or substitutions (see Levenshtein, 1966). Table A.5.1 exemplifies the scoring procedure for a rather complex example.

Table A.5.1
Scoring example, phonetically transcribed.

Target word	k	r	e	m	a			ʝ	ʝ	r	a
Participant’s production	k	r	e	m	a	d	i	ʝ			o
Scoring	C	C	C	C	C	IC (ins)	IC (ins)	C	IC (del)	IC (del)	IC (sub)

Note. C = correct, IC = incorrect; ins = insertion; del = deletion; sub = substitution.

‘Cremadillo’ would be counted as having 6 correct and 5 incorrect phonemes. As explained in the manuscript, these two numbers (6,5) serve as basis for the dependent variable for statistical modelling. Importantly though, we used a binomial probability distribution for statistical modelling, for which word length cannot vary within words. For ‘cremallera’ this means that the number of correct and incorrect phonemes always needs to add up to 9, rather than 11 as in the above example (6+5=11). Hence, we adjusted the counts such that the sum of correct and incorrect phonemes always equaled the original word length. Rescaling was done by multiplying the word length of the target word (in this case 9) by the percentage of correctly produced phonemes (e.g., 9*0.545 = 4.9, rounded off to 5), and subtracting this number from the target word length (e.g., 9-5 = 4, arriving at (5,4) for the cremadillo example). This procedure was taken from de Vos et al. (2018). We refer the reader to the supplementary materials of that paper for a more comprehensive explanation of the issue.

A.6 | Analyses on Raw Naming Latencies

A.6.1 | *Experiment 1*

Table A.6.1.1

Model output for log-transformed raw naming latencies in Experiment 1.

Fixed effects	Estimate	*SE*	*t*	$p(\chi^2)$		
Intercept	7.47	0.05	146.27	**<.001**		
Interference	0.14	0.02	7.17	**<.001**		
Language	0.06	0.09	0.64	.518		
Day	0.15	0.02	6.32	**<.001**		
Interference*Language	0.08	0.04	2.13	**.036**		
Language*Day	0.00	0.05	0.05	.970		
Interference*Day	-0.12	0.04	-3.30	**.001**		
Interference*Language*Day	-0.17	0.08	-2.27	**.024**		
Random effects	**Groups**	**Var**	***SD***	**Corr**		
Item	Intercept	0.06	0.23			
Subject	Intercept	0.08	0.28			
	Interference	0.01	0.09	0.17		
	Day	0.00	0.03	0.31	0.79	
	Int*Day	0.00	0.02	-0.89	0.22	0.15

Note. Significant effects are marked in bold. *SE* = standard error; $p(\chi^2)$ = Chi-square p-value; Var = variance; *SD* = standard deviation; Corr = correlation.

Table A.6.1.2

Model output for log-transformed raw naming latencies split by day in Experiment 1.

	Model output for Day 2				Model output for Day 8			
Fixed effects	Estimate	*SE*	*t*	$p(\chi^2)$	Estimate	*SE*	*t*	$p(\chi^2)$
Intercept	7.40	0.05	140.96	**<.001**	7.54	0.05	143.51	**<.001**
Interference	0.20	0.03	7.98	**<.001**	0.08	0.03	2.72	**.007**
Language	0.05	0.09	0.62	.531	0.06	0.09	0.69	.481
Interference* Language	0.18	0.05	3.64	**<.001**	-0.01	0.06	-0.10	.917
Random effects	Groups	Var	*SD*		Groups	Var	*SD*	
Item	Intercept	0.06	0.25		Intercept	0.05	0.22	
Subject	Intercept	0.08	0.27		Intercept	0.08	0.28	

Note. Significant effects are marked in bold. *SE* = standard error; $p(\chi^2)$ = Chi-square p-value; Var = variance; *SD* = standard deviation; Corr = correlation

Separate t-tests on the log-transformed raw naming latencies per language group on Day 2 revealed that the interference effect is significant in both groups, though more pronounced in the English group ($t(22)=8.40$, $p < .001$; $d = 1.752$) than in the Dutch group ($t(19)=3.28$, $p=.004$, $d = 0.734$) on Day 2.

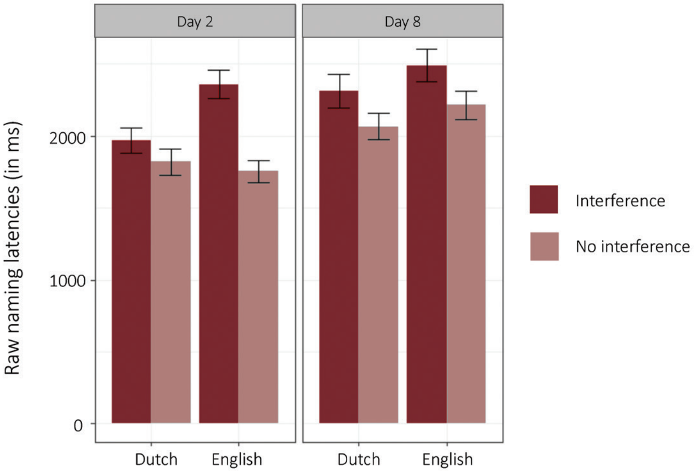

Figure A.6.1

Experiment 1. Raw naming latencies for Spanish productions at final test on day 2 and 8 respectively.

A.6.2 | *Experiment 2*

Table A.6.2

Model outcome for log-transformed raw naming latencies in Experiment 2.

Fixed effects	Estimate	*SE*	*t*	$p(\chi^2)$
Intercept	7.45	0.05	155.24	**<.001**
Interference	0.20	0.02	8.42	**<.001**
Round	0.12	0.10	1.29	.194
Interference * Round	0.09	0.05	1.77	.067
Random effects	**Groups**	**Var**	***SD***	
Item	Intercept	0.04	0.21	
Subject	Intercept	0.08	0.28	

Note. Significant effects are marked in bold. *SE* = standard error; $p(\chi^2)$ = Chi-square p-value; Var = variance; *SD* = standard deviation; Corr = correlation.

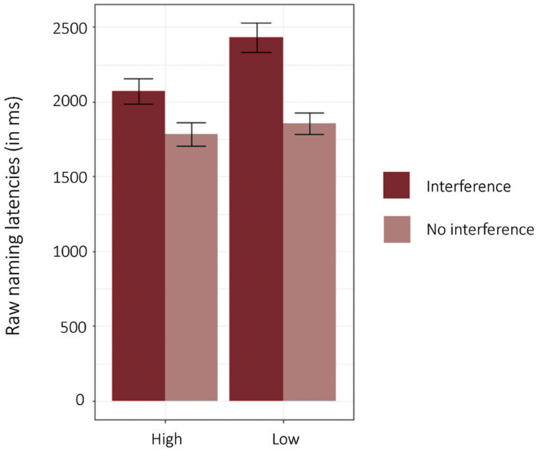

Figure A.6.2

Experiment 2. Raw naming latencies for Spanish productions at final test.

A.7 | Stimulus List for Experiment 2

High frequency items			
English	**Dutch**	**Spanish**	**Item type**
ashtray	asbak	cenicero	Experimental
back	rug	espalda	Experimental
ball	bal	pelota	Experimental
bell	bel	campana	Experimental
belt	riem	cinturón	Experimental
blanket	deken	manta	Experimental
branch	tak	rama	Experimental
bush	struik	arbusto	Experimental
cage	kooi	jaula	Experimental
candle	kaars	vela	Experimental
car	auto	coche	Experimental
carrot	wortel	zanahoria	Experimental
cloud	wolk	nube	Experimental
coat	jas	abrigo	Experimental
dress	jurk	vestido	Experimental
duck	eend	pato	Experimental
egg	ei	huevo	Experimental
fence	hek	alambrada	Experimental
fly	vlieg	mosca	Experimental
gift	cadeau	regalo	Experimental
horse	paard	caballo	Experimental
knife	mes	cuchillo	Experimental
leg	been	pierna	Experimental
lock	slot	cerradura	Experimental
mirror	spiegel	espejo	Experimental
moustache	snor	bigote	Experimental
mug	kopje	taza	Experimental
nose	neus	nariz	Experimental
painting	schilderij	cuadro	Experimental
pig	varken	cerdo	Experimental
pot	pan	olla	Experimental
ring	ring	anhillo	Experimental
steeringwheel	stuur	volante	Experimental
suit	pak	traje	Experimental
suitcase	koffer	maleta	Experimental
table	tafel	mesa	Experimental
wardrobe	kast	armario	Experimental
watch	horloge	reloj	Experimental
window	raam	ventana	Experimental
woman	vrouw	mujer	Experimental
bicycle	fiets	bicicleta	Filler
bird	vogel	pajaro	Filler

English	Dutch	Spanish	Item type
bone	bot	hueso	Filler
book	boek	libro	Filler
bottle	fles	botella	Filler
bowl	kom	cuenco	Filler
bread	brood	pan	Filler
bucket	emmer	cubo	Filler
cheese	kaas	queso	Filler
eye	oog	ojo	Filler
feather	veer	pluma	Filler
garlic	knoflook	ajo	Filler
glasses	bril	gafas	Filler
lid	deksel	tapa	Filler
plate	bord	plato	Filler
sausage	worst	salchicha	Filler
snake	slang	serpiente	Filler
stone	rots	piedra	Filler
tooth	tand	diente	Filler
wood	hout	madera	Filler

Low frequency items

English	Dutch	Spanish	Item type
apron	schort	delantal	Experimental
bagpipe	doedelzak	gaita	Experimental
beak	snavel	pico	Experimental
bib	slab	babero	Experimental
blackberry	braam	mora	Experimental
buckle	gesp	hebilla	Experimental
caterpillar	rups	oruga	Experimental
chalk	krijtje	tiza	Experimental
chisel	beitel	cincel	Experimental
fan	waaier	abanico	Experimental
funnel	trechter	embudo	Experimental
gallows	galg	horca	Experimental
gold bar	staaf	lingote	Experimental
grater	rasp	rallador	Experimental
hinge	scharnier	bisagra	Experimental
iron	strijkijzer	plancha	Experimental
ivy	klimop	yedra	Experimental
marble	knikker	canica	Experimental
mitten	want	manopla	Experimental
mop	dweil	fregona	Experimental
nut	moer	tuerca	Experimental
pacifier	fopspeen	chupete	Experimental
pickle	augurk	pepinillo	Experimental
rake	hark	rastrillo	Experimental

English	Dutch	Spanish	Item type
rattle	rammelaar	sonajero	Experimental
rim	velg	llanta	Experimental
rod	hengel	caña	Experimental
shoehorn	schoenlepel	calzador	Experimental
skate	schaats	patín	Experimental
skunk	stinkdier	mofeta	Experimental
slide	glijbaan	tobogán	Experimental
stork	ooievaar	cigüeña	Experimental
swing	schommel	columpio	Experimental
thimble	vingerhoed	dedal	Experimental
top	tol	peonza	Experimental
water bottle	bidon	cantimplora	Experimental
watering can	gieter	regadera	Experimental
wheelbarrow	kruiwagen	carretilla	Experimental
whip	zweep	látigo	Experimental
whisk	garde	batidor	Experimental
barrier	slagboom	barrera	Filler
coaster	onderzetter	posavasos	Filler
compass	passer	compás	Filler
fire extinguisher	brandblusser	extintor	Filler
hedgehog	egel	erizo	Filler
horseshoe	hoefijzer	herradura	Filler
paintbrush	kwast	brocha	Filler
pawn	pion	peón	Filler
peacock	pauw	pavo	Filler
peel	schil	piel	Filler
peg	wasknijper	pinza	Filler
seesaw	wip	balancín	Filler
shovel	schep	pala	Filler
slingshot	katapult	tirador	Filler
stool	kruk	taburete	Filler
strainer	zeef	colador	Filler
suspenders	bretel	tirantes	Filler
tuning fork	stemvork	diapasón	Filler
whistle	fluit	silbato	Filler
wig	pruik	peluca	Filler

Note. Items are ordered alphabetically in English and per item type.

A.8 | Individual Difference Analyses

A.8.1 | *Cognitive Control*

In exploratory analyses, we asked whether cognitive control ability, as measured in the Simon and Go-NoGo task, predicts forgetting rates in the experiments reported in this manuscript. To that end, we added both Simon task performance (= RT difference incongruent-congruent trials) and Go-NoGo task performance (= false alarm rate) as predictors to the models for both accuracy and reaction times in Experiments 1 and 2. We fit separate models for each of those two predictors and only report relevant interactions including the factor 'Interference' and the respective cognitive control measure.

A.8.1.1 | **Go-NoGo Task Performance**

NoGo false alarm (FA) rate did not modulate interference effects in RTs or accuracy in Experiment 2, and not in RTs in Experiment 1 either (*ps* > .28). In accuracy in Experiment 1, there was a marginal 4-way interaction between FA rate, Language group, Interference and Day (p = .074). Follow-up analyses revealed that FA rate modulated forgetting rates only on Day 8 in the Dutch group ($\beta = 0.68, p = .034$), such that higher FA rates were associated with stronger interference effects. This effect suggests that participants with worse inhibitory control (i.e., higher FA rates), were more affected by language interference, possibly because they were less efficient at applying inhibition to Dutch competitors during the final retrieval in Spanish. It is unclear though why such an effect would only show on Day 8, and only in the Dutch group.

A.8.1.2 | **Simon Task Performance**

Simon task performance did not modulate interference effects in either RTs or accuracy in Experiment 2 (ps > .42). In Experiment 1, performance on the Simon task also did not modulate forgetting rates as measured in RTs (*ps* > .45), but it did modulate forgetting rates in accuracy. Follow-up models clarified that this was only the case on Day 8 and only in the English group: contrary to the NoGo effects reported above, stronger Simon effects were associated with smaller interference effects on Day 8 in the English group ($\beta = 0.30$, $p = .002$). This effect suggests that participants with worse inhibitory control (and hence a stronger Simon effect), were less affected by language interference. Contrary to the conclusion drawn on the basis of the NoGo task performance, this effect suggests that participants with worse inhibitory control were less affected by interference, possibly because they engaged less inhibitory control during the interference phase, and consequently suppressed the Spanish translations less than participants with better inhibitory control (but also stronger interference effects). Again though, it remains puzzling why this effect only emerges on Day 8 and only in one group, and why it is inconsistent with Go-NoGo task performance.

Overall, we urge to take these exploratory results with a grain of salt because our experiments were not set-up to test for individual differences. For a reliable investigation into individual differences a bigger sample size would be called for. What is more, both the Simon and the Go-NoGo task were included as filler tasks and to match participants on cognitive control ability across groups, rather than to predict individual forgetting rates.

A.8.2 | *Multilingual Background*

It is conceivable that participants who have learned multiple foreign languages are more used to dealing with between-language interference and are thus less affected by the experimental manipulation in the current experiments. We tested this by adding the number of previously learned foreign languages as a continuous predictor to the models for accuracy and RTs in both experiments. There were no significant interactions involving this factor (*ps* > .17). Number of languages learned prior to the experiment thus did not modulate forgetting effects either in accuracy or in reaction times in either Experiment 1 or Experiment 2. It should be noted though that our experiments were not designed to test for an effect of multilingual experience on forgetting, both in terms of sample size, and in terms of the distribution of the 'multilingualism' variable here: most participants knew more than two languages (see Figures A.8.1 and A.8.2).

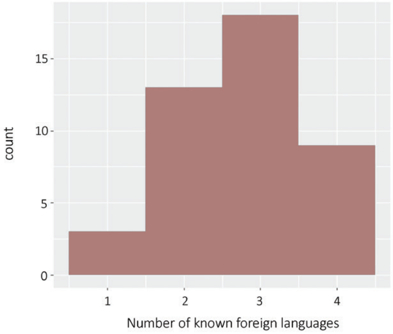

Figure A.8.1
Histogram of the number of foreign languages known by participants in Experiment 1.

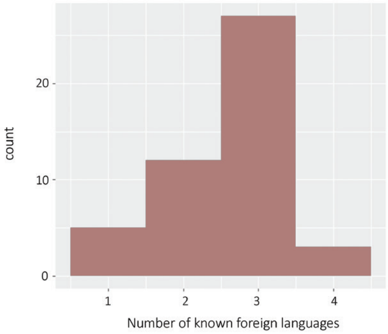

Figure A.8.2
Histogram of the number of foreign languages known by participants in Experiment 2.

A.8.3 | *Age of Onset of Bilingualism*

Research with different types of bilingual populations suggests that early and proficient bilinguals rely less on inhibitory control in speech production than late bilinguals (Costa & Santesteban, 2004). Given this difference, it is possible that late bilinguals suffer more from language-RIF than early bilinguals. We again tested for this by including English AoA as a predictor in the respective statistical models. In Experiment 1, it had no influence on forgetting rates (*ps* > .45). In Experiment 2, however, it did modulate forgetting rates in reaction times ($\beta = 0.08$, $p = .015$, not in accuracy though, $p = .25$) such that participants who started learning English later, showed a stronger interference effect than participants who started learning English earlier on (regardless of which frequency group they were in). It is unclear why the effect was found in Experiment 2 and not in Experiment 1, especially because English was not part of Experiment 2 and thus should have little impact on forgetting. The spread with regard to English AoA, though relatively narrow in general, was similar across the two experiments: most of our participants had started learning English in high school (see Figures A.8.3 and A.8.4).

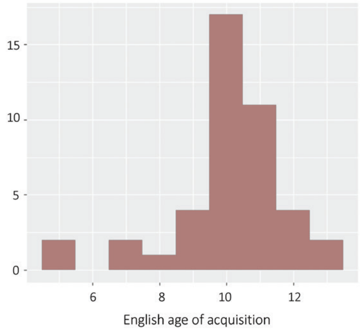

Figure A.8.3
Histogram of age of acquisition of English for participants in Experiment 1.

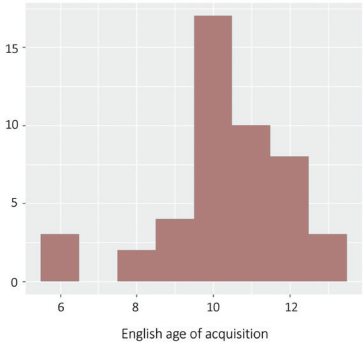

Figure A.8.4
Histogram of age of acquisition of English for participants in Experiment 2.

APPENDIX B

Supplementary materials for Chapter 3

B.1 | Stimulus List

Italian	English	Condition
albero	tree	Set 1
altalena	swing	Set 1
aquilone	kite	Set 1
ascia	axe	Set 1
cannuccia	straw	Set 1
capra	goat	Set 1
cespuglio	bush	Set 1
coltello	knife	Set 1
coperta	blanket	Set 1
cucchiaio	spoon	Set 1
cuffie	headphones	Set 1
foglia	leaf	Set 1
formica	ant	Set 1
freccia	arrow	Set 1
guinzaglio	leash	Set 1
legno	wood	Set 1
lumaca	snail	Set 1
manica	sleeve	Set 1
mattone	brick	Set 1
mosca	fly	Set 1
onda	wave	Set 1
orecchino	earring	Set 1
panchina	bench	Set 1
pipistrello	bat	Set 1
quadro	painting	Set 1
ramo	branch	Set 1
rana	frog	Set 1
scivolo	slide	Set 1
specchio	mirror	Set 1
squalo	shark	Set 1
sughero	cork	Set 1
tenda	curtain	Set 1
baffi	moustache	Set 1
bara	coffin	Set 1
canestro	basket	Set 1
fischietto	whistle	Set 2
ala	wing	Set 2
bambola	doll	Set 2
dado	dice	Set 2
accendino	lighter	Set 2
ago	nail	Set 2
arancia	orange	Set 2
cappello	hat	Set 2
cerniera	zipper	Set 2
chiave	key	Set 2
ciliegia	cherry	Set 2
cintura	belt	Set 2
ciotola	bowl	Set 2
cipolla	onion	Set 2
fiammifero	match	Set 2
frusta	whip	Set 2
gabbia	cage	Set 2
gamba	leg	Set 2
gonna	skirt	Set 2
guscio	shell	Set 2
matita	pencil	Set 2
nuvola	cloud	Set 2
pala	shovel	Set 2
pannolino	diaper	Set 2
pollice	thumb	Set 2
recinto	fence	Set 2
schiena	back	Set 2
scopa	broom	Set 2
semaforo	traffic light	Set 2
spazzola	hairbrush	Set 2
stivale	boot	Set 2
teschio	skull	Set 2
uva	grapes	Set 2
vestaglia	bathrobe	Set 2
zaino	backpack	Set 2
torta	cake	Filler
radice	root	Filler
candela	candle	Filler
sciarpa	scarf	Filler

Italian	English	Condition
aeroplano	airplane	Filler
arco	bow	Filler
guanto	glove	Filler
portafogli	wallet	Filler
impermeabile	raincoat	Filler
pesca	peach	Filler
sedia a rotelle	wheelchair	Filler
rasoio	razor	Filler
tacco	heel	Filler
valigia	suitcase	Filler
dente	tooth	Filler
piatto	plate	Filler
orologio	watch	Filler
bottone	button	Filler
ferro da stiro	iron	Filler
torre	tower	Filler
collana	necklace	Filler
corda	rope	Filler
cravatta	tie	Filler
sega	saw	Filler
tamburo	drum	Filler
reggiseno	bra	Filler
aglio	garlic	Filler
bottiglia	bottle	Filler
fungo	mushroom	Filler
finestra	window	Filler
francobollo	stamp	Filler
coperchio	lid	Filler
osso	bone	Filler
ponte	bridge	Filler
completo da uomo	suit	Filler

APPENDIX C

Supplementary materials for Chapter 4

C.1 | Pre-Test Item List Experiment 1

Spanish	English	Set	Spanish	English	Set
plancha	iron	1	salchicha	sausage	2
tapa	lid	1	tiburón	shark	2
taladro	drill	1	hacha	axe	2
cartera	wallet	1	fregona	mop	2
cuenco	bowl	1	columpio	swing	2
rana	frog	1	toro	bull	2
pavo	turkey	1	mechero	lighter	2
tiza	chalk	1	ciervo	deer	2
sujetador	bra	1	pala	shovel	2
látigo	whip	1	cuchillo	knife	3
ajo	garlic	1	lata	can	3
caracol	snail	1	coche	car	3
ardilla	squirrel	1	ventana	window	3
muñeca	doll	1	espalda	back	3
enfermera	nurse	1	manta	blanket	3
hoja	leaf	1	cuchara	spoon	3
ladrillo	brick	1	almohada	pillow	3
silbato	whistle	1	mofeta	skunk	3
hueso	bone	1	corbata	tie	3
dado	dice	1	serpiente	snake	3
canasta	basket	1	pájaro	bird	3
guante	glove	1	cadena	chain	3
bufanda	scarf	1	reloj	watch	3
bandeja	tray	2	rayo	lightning	3
escoba	broom	2	cebolla	onion	3
bolsillo	pocket	2	cuadro	painting	3
cubo	bucket	2	manga	sleeve	3
cremallera	zipper	2	rama	branch	3
regla	ruler	2	perro	dog	3
hongo	mushroom	2	soga	rope	3
alambrada	fence	2	abrigo	coat	3
hormiga	ant	2	falda	skirt	3
pulsera	bracelet	2	mono	monkey	3
ataúd	coffin	2	taza	mug	3
melocotón	peach	2	jaula	cage	3
foca	seal	2	árbol	tree	3
silla	chair	2	cerradura	lock	3

Spanish	English	Set
bigote	moustache	3
loro	parrot	3
ola	wave	3
ala	wing	3
paja	straw	3
flecha	arrow	3
pierna	leg	3
cerillo	match	3
cicatriz	scar	3
caballo	horse	3
raíz	root	3
cigüeña	stork	3
bala	bullet	3
vela	candle	3
portería	goal	3
nube	cloud	3
pecho	chest	3
madera	wood	3
tacón	heel	3
burro	donkey	3
traje	suit	3
mapache	raccoon	3
camarero	waiter	3
llave	key	3
paloma	pigeon	3
calavera	skull	3
lápiz	pencil	3
cinturón	belt	3
pato	duck	3
vestido	dress	3

C.2 | Item Selection

If a participant did not know one or more words from the ideal base-set (i.e., the first 46 words in the English pre-test), these had to be replaced with known words from the remaining pre-test items. A Matlab script, adapted from Chapter 2, took care of the replacement procedure. In general, the script replaced items one after the other, and replacements were chosen such that they were as close to the respective base-set item as possible in terms of word length, phonological make-up and semantic similarity. To make sure that this was the case, the script initially chose the word that was semantically closest to the base-set item that needed to be replaced. Following the procedure in Chapter 2, the script chose the replacement option with the smallest semantic distance to the to-be-replaced word (see Appendix A.2 for details).

Subsequently, also similar to Chapter 2, the script checked whether this replacement option was within +-1 syllable length in Spanish from the to-be-replaced word, whether its Spanish and English Levenshtein distance with other words in the entire set was at least 2 and whether its semantic distance to words in the other subset would not exceed a predefined threshold of 0.68. When the to-be-replaced word was a compound in Dutch, the script would first attempt to find a replacement that was also a compound. When any of the above conditions were not met by the replacement option, it was discarded and a new replacement option was selected and evaluated. This was repeated until a viable replacement candidate was found.

When a replacement had been found, it was added to its prospective subset (i.e., the subset that the base set item it was supposed to replace belonged to) and the following three criteria were assessed (the first two are identical to Chapter 2; see Appendix A.2 for details):

1 Word length in Spanish and English (in syllables): the average word length for the two subsets had to be comparable (i.e., not statistically different);
2 Semantic similarity *within* subsets: the mean semantic similarity between items within each set could not be statistically different;
3 Frequency in Dutch (log lemma frequencies): the average Dutch log frequency had to be roughly the same in the two subsets (i.e., not statistically different).

For 23 out of the 31 tested participants in Experiment 1 and for 29 out of 86 tested participants in Experiment 2, the script could not find replacements for some of the unknown words from the base set. In those cases, we ran a more lenient script for item selection. In a first try, this more lenient script dropped the initial across-set semantic similarity constraint (successful for 15 participants in Experiment 1 and for 25 in Experiment 2). If this did not help, the within-set semantic similarity constraint was also dropped (successful for three additional participants in Experiment 1 and for four additional participants in Experiment 2). In Experiment 1, for four participants the script failed even after the second adjustment, these participants were sent home. In Experiment 2 this happened for five people in total. In order to avoid sending those participants home, we instead counted one of the unknown words (the one where the script failed to find a replacement) as known and reran the script from scratch. The script then succeeded with all criteria for two participants, for the remaining three it succeeded after dropping the across-set semantic similarity criterion. The unknown words that were counted as known were later excluded from analysis. The resulting item set characteristics are summarized in Table 4.2.

C.3 | Apparatus and Task Details

C.3.1 | *Apparatus*

All tasks were administered on a Dell T3610 computer (3,7Ghz Intel Quad Core, 8GB RAM), running Windows 7 and using the stimulus presentation software Presentation (Version 19.0, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com). The computer screen (BenQ XL 2420Z, 24-inch) was set to white, with a resolution of 1920 x 1080 pixels at a refresh rate of 60 Hz. All audio stimuli were presented to the participants via headphones (Sennheiser HD201), and all oral responses were recorded via a microphone (Shure SM-57) in WAV format using a Behringer X-Air XR18 digital mixer.

C.3.2 | *Task Timings*

C.3.2.1 | English Pre-Test and Post-Test

Each trial started with a 500 ms blank screen, followed by the presentation of the picture in the center of the screen. There was no time limit for participants to provide their answer. The experimenter coded the answers for correctness by pressing one of two keys: ENTER for incorrect and SPACE for correct. There was no feedback; the next trial started immediately after the experimenter’s button press.

C.3.2.2 | Learning Tasks

Stimulus timings in all learning tasks were identical to those reported in Chapter 2 (see Appendix A.4).

C.4 | Raw Naming Latencies from English Pre- and Post-test

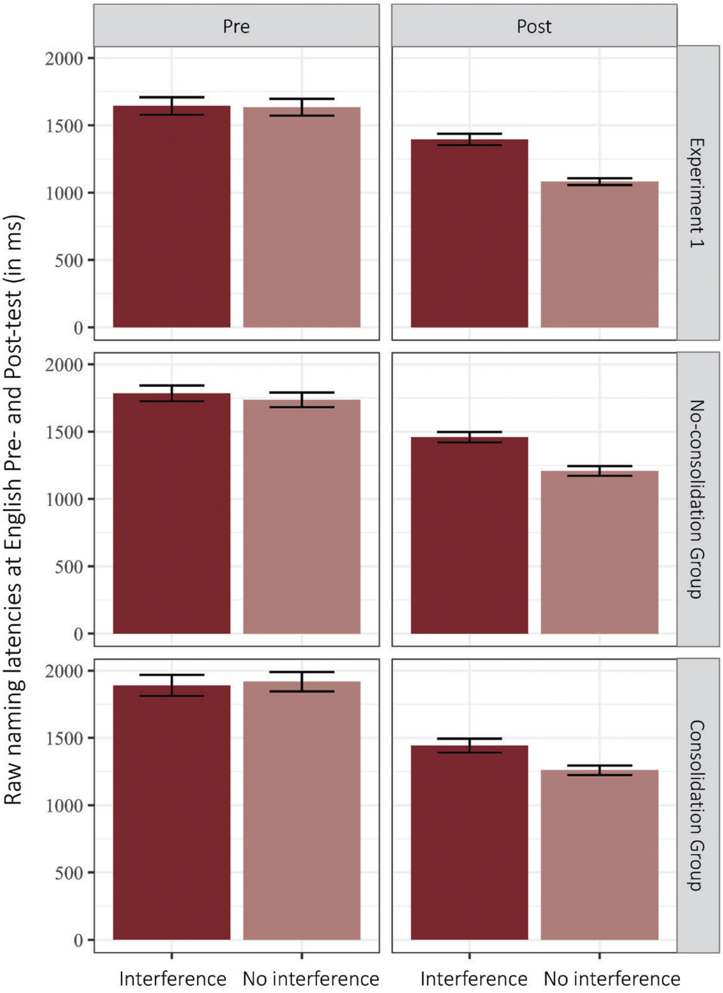

Figure C.4
Raw naming latencies (in ms) from English pre- and post-tests in Experiment 1 and the two groups from Experiment 2, split by interference condition.

C.5 | Pre-Test Item List Experiment 2

Spanish	English	Set
alambrada	fence	1
almohada	pillow	1
arbusto	bush	1
baúl	chest	1
burro	donkey	1
canasta	basket	1
cartera	wallet	1
cerillo	match	1
dado	dice	1
enchufe	socket	1
flecha	arrow	1
grifo	tap	1
hormiga	ant	1
ladrillo	brick	1
loro	parrot	1
ola	wave	1
pepinillo	pickle	1
rama	branch	1
regla	ruler	1
sello	stamp	1
silbato	whistle	1
taza	mug	1
ventana	window	1
aguja	needle	2
bigote	moustache	2
columpio	swing	2
concha	shell	2
cubo	bucket	2
cuna	cradle	2
escarabajo	beetle	2
escoba	broom	2
gorro	hat	2
grapadora	stapler	2
hueso	bone	2
jabón	soap	2
lata	can	2
manga	sleeve	2
marco	frame	2
moneda	coin	2
nube	cloud	2
pala	shovel	2

Spanish	English	Set
pato	duck	2
salchicha	sausage	2
taburete	stool	2
tarro	jar	2
tobogán	slide	2
abanico	fan	3
abrigo	coat	3
ajo	garlic	3
ala	wing	3
araña	spider	3
árbol	tree	3
arco	bow	3
ardilla	squirrel	3
avestruz	ostrich	3
bala	bullet	3
bastón	cane	3
batidor	whisk	3
bufanda	scarf	3
caballo	horse	3
cacahuete	peanut	3
cadena	chain	3
caja	box	3
caña	rod	3
canica	marble	3
caracol	snail	3
cebolla	onion	3
cerdo	pig	3
cinturón	belt	3
clavo	nail	3
coche	car	3
colador	strainer	3
corbata	tie	3
cremallera	zipper	3
cuchara	spoon	3
cuchillo	knife	3
cuenco	bowl	3
diente	tooth	3
embudo	funnel	3
espejo	mirror	3
falda	skirt	3
fregona	mop	3

Spanish	English	Set
guante	glove	3
hacha	axe	3
hoja	leaf	3
huevo	egg	3
jaula	cage	3
lápiz	pencil	3
llanta	rim	3
llave	key	3
machacador	masher	3
madera	wood	3
manopla	mitten	3
manta	blanket	3
mapache	raccoon	3
mechero	lighter	3
melocotón	peach	3
mono	monkey	3
mosca	fly	3
muñeca	doll	3
naranja	orange	3
oruga	caterpillar	3
paja	straw	3
pájaro	bird	3
paloma	pigeon	3
pañal	diaper	3
paraguas	umbrella	3
pastel	cake	3
pavo	turkey	3
peluca	wig	3
pierna	leg	3
plancha	iron	3
pluma	feather	3
pulsera	bracelet	3
puño	fist	3
queso	cheese	3
raíz	root	3
rallador	grater	3
rana	frog	3
regalo	gift	3
reloj	watch	3
serpiente	snake	3
sierra	saw	3
silla	chair	3
soga	rope	3
sujetador	bra	3
taladro	drill	3
tapa	lid	3
tiburón	shark	3
tiza	chalk	3
toalla	towel	3
tornillo	screw	3
tortuga	turtle	3
traje	suit	3
tuerca	nut	3
uvas	grape	3
vela	candle	3
vestido	dress	3
yedra	ivy	3
zanahoria	carrot	3

APPENDIX D

Supplementary materials for Chapter 5

D.1 | Participant Recruitment and Exclusion Criteria

To be invited for participation in the experiment, participants had to fulfill a number of criteria: German had to be their only mother tongue (which disqualified seven sign-ups out of the 481 sign-ups we received) and their planned study abroad had to stretch a minimum of four and a maximum of seven months (which disqualified 64 people). We also excluded participants who planned to go abroad at a later point in time (starting January or February 2019, disqualifying ten people), as well as people who at the time of sign-up had already been in Spain for more than two weeks (getting a proper baseline was no longer possible for those cases, resulting in 74 disqualifications). The remaining 326 were asked to fill in a short questionnaire about their planned university courses while abroad. We were interested in whether they were planning to take courses in English or Spanish. People who indicated planning to take more than 50% of their courses in English were not invited (N = 32) and those who did not fill in this survey were also not invited (N = 100).

On the basis of these exclusions, we ended up inviting 194 German native speakers. 34 of those did not finish the tasks of the first session (T1) and were hence excluded from the remainder of the experiment. An additional eight did complete T1 tasks, but had empty recordings and were hence also excluded from continuing with the experiment. The remaining 152 participants proceeded to the second session (T2). Out of those, four decided to prolong their study abroad and hence had to leave the study after the T2 measurement. Seven did not complete the T2 tasks and eight had empty recordings and were hence excluded from the rest of the experiment as well. The remaining 134 participants moved on to the last session of the experiment (T3). Twelve of them did not complete the T3 tasks and five had empty recordings, resulting in a pre-final dataset of 117 participants, for whom we have complete Spanish recordings for all three time points. Out of those, another 18 only have partial fluency data for T2 and T3, resulting in a final data set of 99 participants, for whom we have complete T2 and T3 data. Next to exclusions based on data availability, we also excluded two participants because their audio recordings indicated that, prior to their spoken response, they had typed on their computer keyboards on more than 20% of the trials of the Spanish naming task at either T2 or T3, casting doubt on the integrity of their answers.

D.2 | Email Reminders

We repeatedly checked whether participants had completed their tasks, and if not, sent them reminders. The first reminder was sent four to six days after initial invitation (necessary for 46% of participants at T2 and 71% at T3), a second reminder after eight to ten days (necessary for 18% at T2 and 39% at T3), a third reminder after 12 to 14 days (necessary only at T3 for 2%), and in extreme cases one final reminder after 20 days (sent to 2% of participants at T3).

D.3 | Session Scheduling

D.3.1 | *T2 Timing*

As described in the manuscript, we rescheduled the T2 session for participants who went to Germany for Christmas. Participants who went back to Germany for a week or longer and who were going to have less than three times as many days left in Spain after Christmas as they went to Germany for were invited to complete T2 before Christmas (29 out of the 97 participants). The rest of the participants were invited for T2 two weeks before they left Spain.

D.3.2 | *T3 Timing*

As described in the manuscript, T3 was timed such that participants had either not been abroad since their return to Germany or only briefly (and at least one month ago). 16 out of our 97 participants indicated to have been to Spanish-speaking countries between T2 and T3 (M = 20.5 days, SD = 18 days, range = 4-60 days).

D.4 | Stimulus List for Spanish Vocabulary Test

English	Spanish	German	Item Type	Cognate (0=non-cognate 1= Ger-Eng-Spa 2= Eng-Spa)	T2 order	T3 order
strawberry	fresa	Erdbeere	Practice	0	4	2
tree	arból	Baum	Practice	0	2	3
dog	perro	Hund	Practice	0	3	1
car	coche	Auto	Practice	0	1	4
chocolate	chocolate	Schokolade	Test	1	75	17
doll	muñeca	Puppe	Test	0	22	83
bike	bicicleta	Fahrrad	Test	2	97	25
chair	silla	Stuhl	Test	0	104	57
telephone	telefono	Telefon	Test	1	91	37
shoe	zapato	Schuh	Test	0	111	134
butterfly	mariposa	Schmetterling	Test	0	5	46
bus	bus	Bus	Test	1	68	48
shark	tiburón	Hai	Test	0	49	59
lightbulb	bombilla	Glühbirne	Test	0	126	79
table	mesa	Tisch	Test	0	98	69
bottle	botella	Flasche	Test	2	7	12
skirt	falda	Rock	Test	0	46	75
fingernail	uña	Fingernagel	Test	0	62	52
suit	traje	Anzug	Test	0	85	32
lemon	limón	Zitrone	Test	2	125	81
glove	guante	Handschuh	Test	0	51	125
can	lata	Blechdose	Test	0	115	114
watch	reloj	Armbanduhr	Test	0	102	54
pearl	perla	Perle	Test	1	20	76
bellpepper	pimiento	Paprika	Test	0	106	102
garlic	ajo	Knoblauch	Test	0	32	123
sleeve	manga	Ärmel	Test	0	93	35
circle	circulo	Kreis	Test	2	28	124
crutch	muleta	Krücke	Test	0	13	15
heel	tacón	(Schuh)absatz	Test	0	120	126
pan	sartén	Pfanne	Test	0	59	24
pharmacy	farmacia	Apotheke	Test	2	71	21
sausage	salchicha	Wurst	Test	0	21	47
spoon	cuchara	Löffel	Test	0	110	97
screw	tornillo	Schraube	Test	0	108	133

English	Spanish	German	Item Type	Cognate (0=non-cognate 1= Ger-Eng-Spa 2= Eng-Spa)	T2 order	T3 order
tomato	tomate	Tomate	Test	1	25	88
key	llave	Schlüssel	Test	0	38	36
brocoli	brócoli	Brokkoli	Test	1	121	132
duck	pato	Ente	Test	0	135	140
flower	flor	Blume	Test	2	80	142
knife	cuchillo	Messer	Test	0	48	19
map	mapa	(Land)karte	Test	2	79	113
apron	delantal	Schürze	Test	0	74	121
horse	caballo	Pferd	Test	0	42	14
fork	tenedor	Gabel	Test	0	47	63
castle	castillo	Burg	Test	2	94	95
umbrella	paraguas	Regenschirm	Test	0	89	11
suitcase	maleta	Koffer	Test	0	131	9
straw	paja	Strohhalm	Test	0	127	109
spider	araña	Spinne	Test	0	29	86
sofa	sofá	Sofa	Test	1	30	73
fan	abanico	Fächer	Test	0	87	115
waiter	camarero	Kellner	Test	0	45	30
button	botón	Knopf	Test	2	67	51
pigeon	paloma	Taube	Test	0	84	8
orange	naranja	Orange	Test	0	70	70
socket	enchufe	Steckdose	Test	0	8	94
salmon	salmón	Lachs	Test	2	90	90
scar	cicatriz	Narbe	Test	0	39	141
keyboard	teclado	Tastatur	Test	0	12	6
plate	plato	Teller	Test	2	37	49
ring	anillo	Ring	Test	0	54	77
snail	caracol	Schnecke	Test	0	134	129
melon	melón	Melone	Test	1	114	107
candle	vela	Kerze	Test	0	117	41
foot	pie	Fuss	Test	0	10	89
lake	lago	See	Test	2	60	31
goat	cabra	Ziege	Test	0	11	50
hairbrush	cepillo	Haarbürste	Test	0	124	80
lettuce	lechuga	Salat	Test	0	56	22
kangaroo	canguro	Känguru	Test	1	81	139
lock	candado	Schloss	Test	0	105	100

English	Spanish	German	Item Type	Cognate (0=non-cognate 1= Ger-Eng-Spa 2= Eng-Spa)	T2 order	T3 order
calculator	calculador	Taschenrechner	Test	2	35	93
blanket	manta	Decke	Test	0	128	120
gift	regalo	Geschenk	Test	0	64	34
cable	cable	Kabel	Test	1	63	99
wave	ola	Welle	Test	0	69	78
scarf	bufanda	Schal	Test	0	138	111
wardrobe	armario	Schrank	Test	0	122	42
envelope	sobre	Briefumschlag	Test	0	17	127
tap	grifo	Wasserhahn	Test	0	34	117
elephant	elefante	Elefant	Test	1	142	5
rug	alfombra	Teppich	Test	0	119	38
shell	concha	Muschel	Test	0	118	106
needle	aguja	Nadel	Test	0	31	64
lion	león	Löwe	Test	1	72	10
bellybutton	ombligo	Bauchnabel	Test	0	41	7
broom	escoba	Besen	Test	0	23	16
onion	cebolla	Zwiebel	Test	0	143	84
cross	cruz	Kreuz	Test	1	113	29
egg	huevo	Ei	Test	0	86	28
screen	pantalla	Bildschirm	Test	0	50	55
cow	vaca	Kuh	Test	0	44	66
iron	plancha	Bügeleisen	Test	0	58	119
sandal	sandalia	Sandale	Test	1	36	62
cloud	nube	Wolke	Test	0	73	105
bridge	puente	Brücke	Test	0	136	131
soap	jabón	Seife	Test	0	76	144
whale	ballena	Wal	Test	0	140	61
pig	cerdo	Schwein	Test	0	116	122
boot	bota	Stiefel	Test	2	16	82
bone	hueso	Knochen	Test	0	123	74
leg	pierna	Bein	Test	0	99	20
radio	radio	Radio	Test	1	137	87
earring	pendiente	Ohrring	Test	0	78	33
hospital	hospital	Krankenhaus	Test	2	144	44
peanut	cacahuete	Erdnuss	Test	0	40	45
feather	pluma	Feder	Test	0	53	138
seagull	gaviota	Möwe	Test	0	55	116

English	Spanish	German	Item Type	Cognate (0=non-cognate 1= Ger-Eng-Spa 2= Eng-Spa)	T2 order	T3 order
Vacuum cleaner	aspirador/a	Staubsauger	Test	0	18	104
ashtray	cenicero	Aschenbecher	Test	0	109	136
nose	naríz	Nase	Test	0	112	27
potato	patata	Kartoffel	Test	2	96	137
cage	jaula	Käfig	Test	0	77	128
pineapple	piña	Ananas	Test	0	83	43
goal	gol	Tor	Test	2	130	92
carrot	zanahoria	Möhre	Test	0	133	56
mirror	espejo	Spiegel	Test	0	141	96
lamp	lámpara	Lampe	Test	1	100	68
stamp	sello	Briefmarke	Test	0	61	135
monkey	mono	Affe	Test	0	95	85
truck	camión	Lastwagen	Test	0	103	65
salt	sal	Salz	Test	1	88	40
bracelet	pulsera	Armband	Test	0	101	91
pencil	lápiz	Bleistift	Test	0	15	13
island	isla	Insel	Test	1	66	53
dress	vestido	Kleid	Test	0	65	23
backpack	mochila	Rucksack	Test	0	52	39
eyebrow	ceja	Augenbraue	Test	0	26	112
frog	rana	Frosch	Test	0	82	98
hand	mano	Hand	Test	0	107	110
bear	oso	Bär	Test	0	139	67
train	tren	Zug	Test	2	9	60
pacifier	chupete	Schnuller	Test	0	92	71
pumpkin	calabaza	Kürbis	Test	0	129	18
pillow	almohada	Kissen	Test	0	132	101
banana	banana	Banane	Test	1	24	58
trafficlight	semáforo	Ampel	Test	0	27	108
eggplant	berenjena	Aubergine	Test	0	6	26
battery	batería	Batterie	Test	1	14	103
belt	cinturón	Gürtel	Test	0	43	130
mountain	montaña	Berg	Test	2	19	118
blackberry	mora	Brombeere	Test	0	57	143
cauliflower	coliflor	Blumenkohl	Test	2	33	72

Note. Items are presented in T1 order. For T2 and T3 order see the respective columns.

D.5 | Model Comparisons

Table D.5

Model comparisons between the base model (with only Session as predictor) and the initial predictor models (with Session*Predictor interactions).

Model	AIC	BIC	χ^2	$p(\chi^2)$
Baseline model				
~ Sess	109664.85	109697.61	*NA*	*NA*
Participant-level predictors				
~ Sess * Spanish frequency of use	109393.41	109442.53	275.45	**<.001**
~ Sess * English/German ratio	109667.49	109716.61	1.37	.505
~ Sess * Amount Spanish experience	109595.00	109644.00	73.62	**<.001**
~ Sess * German letter fluency	109662.30	109711.43	6.55	**.038**
~ Sess * English letter fluency	109633.99	109683.12	34.87	**<.001**
~ Sess * Germany category fluency	109664.73	109713.86	4.12	.127
~ Sess * English category fluency	109659.51	109708.64	9.34	**.009**
~ Sess * Integrative motivation	109625.60	109674.73	43.26	<.001*
~ Sess * Anxiety	109620.26	109669.38	48.60	<.001*
~ Sess * Instrumental motivation	109587.20	109636.33	81.65	**<.001**
~ Sess * Attrition judgment	109559.07	109608.19	109.79	**<.001**
~ Sess * LTM capacity	109667.86	109716.99	0.99	.608
~ Sess * T2 performance	109312.79	109361.91	356.07	**<.001**
Item-level predictors				
~ Sess * Spanish log frequency	109619.90	109669.00	49.00	**<.001**
~ Sess * German log frequency	109636.45	109685.58	32.41	<.001†
~ Sess * Cognate status	109627.19	109676.32	41.67	**<.001**

Note. Highlighted rows reflect the predictors that significantly improved model fit compared to the base model with only Session as predictor. The highlighted rows hence reflect predictors that entered the final full model. * We chose one of the three motivation subscores for the final model despite the fact that they all improved model fit. We selected instrumental motivation because it has the lowest AIC and BIC values out of the three. † Both German and Spanish log frequency counts were predictive of forgetting, yet in a direct comparison, Spanish log frequency was the better predictor ($p < .001$).

D.6 | Fluency Task Details

For the letter fluency tasks, for each of the two languages, we selected three frequent word onset letters and randomly assigned them to the three sessions with the constraint that no letter that was part of the English fluency tasks would appear in the German fluency tasks and vice versa. For the semantic category fluency tasks, we compiled a list of 18 possible categories based in part on semantic category norms (Battig & Montague, 1969; Ruts et al., 2004; Overschelde et al., 2004). We then pre-tested those with seven participants in German and English, chose the 12 categories with the most answers, and assigned each category to a language and session, choosing the easiest for the English tasks to make sure participants would be able to do the task even if their English proficiency was low.

Table D.6

Letters and categories chosen for the fluency tests.

	English		German	
	Letter	Categories	Letter	Categories
T1	F (2510, 3126)	land animals, professions	P (2337, 4605)	kitchen utensils, vegetables
T2	A (2720, 3649)	clothes, fruit	M (2212, 5264)	hygiene/bathroom supplies, transportation means
T3	D (3026, 4175)	body parts, electronic devices	B (2820, 8338)	sports, office supplies

Note. Numbers in brackets reflect an estimate of the number of words that start with the letter in question in English or German respectively (counts are based on English and German lemmas in the Celex database, followed by counts based on the English and German Subtlex databases).

D.7 | Complete Questionnaires at T1, T2 and T3

D.7.1 | *T1-specific Questions*

Wie alt bist du? [How old are you?]

Geschlecht [gender]
□ Männlich [male]
□ Weiblich [female]
□ Sonstiges [other]

In welcher Stadt oder Region bist du aufgewachsen? Wenn du in mehreren Regionen aufgewachsen bist, nenne sie bitte alle chronologisch und mit einer ungefähren Zeitangabe, wie lange du im jeweiligen Gebiet gewohnt hast. [In which city or region did you grow up? If you grew up in multiple regions, please name them all together with a rough time indication of when you lived in the respective region.]

In welcher Sprache war deine schulische Ausbildung? Es geht um nicht-sprachliche Fächer (Mathe, Physik, Geschichte etc.). Mehrere Antworten sind möglich: gib dann bitte für jede Sprache an, welche Fächer in dieser Sprache unterrichtet wurden. Wenn alles auf Deutsch lief, schreibe dann einfach 'alle' in das Feld neben 'Deutsch'. [In which language was your school education? This concerns non-language subjects, such as math, physics and history. Please indicate for each language, which subjects were taught in it. If all courses were taught in German, just enter 'all' in the field next to 'German'.]
□ Deutsch [German]
□ Englisch [English]
□ Französisch [French]
□ Spanisch [Spanish]
□ Sonstiges [other]:

Bist du momentan Student? [Are you currently a student?]
□ Ja [yes]
□ Nein [no]

Wenn ja, an welcher deutschen Uni studierst du? [If so, at which German university do you study?]

In welchem Studienjahr befindest du dich? [In which study year are you?]
□ 1. Jahr Bachelor [1st year Bachelor]
□ 2. Jahr Bachelor [2nd year Bachelor]

□ 3. Jahr Bachelor [3rd year Bachelor]
□ 1. Jahr Master [1st year Bachelor]
□ 2. Jahr Master [2nd year Bachelor]
□ Sonstiges: [other]

In welcher Sprache studierst du (in Deutschland)? Mehrere Antworten sind möglich. Gib für jede Sprache bitte an, zu wie viel Prozent dein Studium in dieser Sprache stattfindet. [Which language do you study in? Multiple answers are possible. For each language, please indicate to what percent your studies are held in that language.]
□ Deutsch [German]
□ Englisch [English]
□ Französisch [French]
□ Spanisch [Spanish]
□ Sonstiges [other]

Wann bist du nach Spanien umgezogen (oder wann wirst du umziehen)? [When did you move to Spain (or when will you move to Spain)?]

Während deines Auslandsaufenthaltes wirst du: [During your study abroad you will:]
□ An der Gastuni studieren [study at the host university]
□ Ein Praktikum absolvieren [do an internship]
□ Als Au-pair arbeiten [work as au-pair]
□ Anderweitig arbeiten [work elsewhere]
□ Sonstiges [other]

Warst du bisher schon einmal länger als 3 Monate im spanischsprachigen Ausland? [Have you ever lived in a Spanish-speaking country for more than 3 months?]
□ Ja [yes]
□ Nein [no]

Wenn ja, wo und wie lange warst du im spanischsprachigen Ausland? [If so, where and for how long did you live there?]

Warst du bisher schon einmal länger als 3 Monate im nicht-spanischsprachigen Ausland? [Have you ever lived in a non-Spanish-speaking country for more than 3 months?]
□ Ja [yes]
□ Nein [no]

Wenn ja, wo und wie lange warst du im nicht-spanischsprachigen Ausland? [If so, where and for how long did you live there?]

Welche Fremdsprachen sprichst du? [Which foreign languages do you speak?]
□ Englisch [English]
□ Französisch [French]
□ Spanisch [Spanish]
□ Niederländisch [Dutch]
□ Portugiesisch [Portuguese]
□ Italienisch [Italian]
□ Russisch [Russian]
□ Chinesisch [Chinese]
□ Arabisch [Arabic]
□ Katalanisch [Catalan]
□ Latein [Latin]
□ Sonstiges [other]

Beschreibe auf Spanisch, warum du dich für einen Auslandsaufenthalt in Spanien entschieden hast. [Describe in Spanish why you chose to study abroad in Spain.]

Hast du dich auf den Auslandsaufenthalt vorbereitet? [Did you prepare for the study abroad?]
□ Ja, mit einem Sprachkurs [yes, with a language course]
□ Ja, mit Büchern über Land und Leute [yes, with books about the country and people]
□ Ja, durch einen online Kurs, bzw. eine Sprachlernapp [yes, via an online language course / a language learning app]
□ Ja, administrative (Immatrikulation an der Gastuni, Flugtickets etc.) [yes, administratively (enrolled at host university, flight tickets, etc.)]
□ Nein, ich kenne das Land und die Sprache bereits gut [no, I already know the country and the language well]
□ Nein, keine Zeit [no, no time]
□ Nein, kein Interesse [no, no interest]
□ Sonstiges: [other]

Was erhoffst du dir von deinem Auslandsaufenthalt? [What do you expect from your study abroad?]
□ Spanisch zu lernen / zu perfektionieren [learn / perfectionate Spanish]
□ Spanische Freundschaften [Spanish friends]
□ International Freundschaften [international friends]

□ Einfach mal raus aus Deutschland [just to get out of Germany once]
□ Neue Kultur kennen lernen [getting to know a new culture]
□ Sommer, Sonne, Strand … [summer, sun, beach …]
□ Keine Erwartungen [no expectations]
□ Sonstiges [other]

Hast du dich bewusst für Spanien entschieden? [Did you specifically choose Spain as your study abroad destination?]
□ Ja, definitiv, wenn ein Auslandsjahr, dann im spanischsprachigen Raum [yes, if I go abroad, then to a Spanish-speaking country]
□ Ja, aber ich wäre auch woanders hingegangen [yes, but I would have also gone somewhere else]
□ Nein, Hauptsache ins Ausland [no, any country abroad would have been good]
□ Nein, ich wollte eigentlich woanders hin [no, I actually wanted to go somewhere else]
□ Nein, aber in Spanien gab es noch genügend Plätze [no, but in Spain there were still enough spots]
□ Sonstiges [other]

D.7.2 | *T2-specific Questions*

Beschreibe auf Spanisch, was dir an deinem Auslandsaufenthalt besonders gut und was dir weniger gut gefallen hat. Würdest du einen solchen Auslandsaufenthalt weiterempfehlen? [Describe in Spanish what you liked most and least about your study abroad. Would you recommend a study abroad to other people?]

Wie würdest du deinen Auslandsaufenthalt in den folgenden Aspekten bewerten? [How would you describe your study abroad in the following aspects?]

1 sehr schlecht / total unzufrieden [very bad, completely unsatisfied] – 4 neutral – 7 sehr gut, vollkommen zufrieden [very good, completely satisfied]

□ Insgesamt [overall]
□ Aus sprachlicher Sicht (Hast du sprachlich Fortschritte gemacht? Kann sich auf Spanish aber auch Englisch beziehen?) [Linguistically (Did you improve your language skills, either English or Spanish?)]
□ Aus sozialer Sicht (Warst du zufrienden mit deinem Sozialleben, oder hättest du lieber mehr / weniger Kontakte geknüpft?) [Socially (Were you happy with your social life abroad, or would you have liked to have more social contacts?)]
□ Aus fachlicher Sicht (Hast du viel gelernt? War es für dein Fachstudium bereichernd?) [Professionally (Did you learn a lot? Was it beneficial for your studies?]

Was beschreibt deine Wohnungssituation im Ausland am besten? [What best describes your living situation abroad?]
□ WG / Wohnheim [shared living arrangement]
□ Gastfamilie [host family]
□ Eigene Wohnung [own apartment]
□ Other:

Welche Kurse hast du an der spanischen/lateinamerikanischen Uni besucht? Trage jeweils den Namen des Kurses ein, die Sprache in der er abgehalten wurde und wie viele Stunden pro Woche er lief. [Which courses did you follow at the Spanish university? Enter the name of the course, the language it was held in and how many hours it took per week.]

Wie integriert würdest du sagen warst / bist du in die spanische Gemeinschaft / Kultur und Sprache? [How integrated were you in the Spanish culture and language?]

1 überhaupt nicht integriert [not integrated at all] – 4 neutral – 7 vollständig integriert [completely integrated]

- Sprache [language]
- Kultur / Gesellschaft [culture]

Hast du Spanien/Lateinamerika während deines Auslandsaufenthaltes verlassen? Uns geht es hier um Reisen / kurze Trips in andere Länder, wo nicht Spanisch gesprochen wird, also zum Beispiel, wenn du einen Freund in einem anderen nicht-spanischsprachigem Land besucht hast. [Did you leave Spain during your study abroad? We specifically mean short trips to other countries, for example, to other countries where no Spanish is spoken.]
□ Ja [yes]
□ Nein [no]

Wenn ja, gib für alle Reisen in nicht-spanischsprachige Gebiete an, wohin und für wie lange du dort warst. [If so, indicate for all trips, where you went and for how long you went.]

Bist du nach Ende des Semesters noch privat in Spanien geblieben, oder wirst du noch bleiben? [Did you privately stay in Spain after the end of your study abroad semester?]
□ Ja [yes]
□ Nein [no]

Wenn ja, wie viele Wochen bist du noch geblieben, bzw. wirst du noch bleiben? [If so, how many extra weeks did or will you stay?]

Wenn ja, hat sich in der Zeit dein Sprachgebrauch geändert, bzw. wird er sich wohlmöglich ändern? Bist du zum Beispiel mit deutschen Freunden gereist, und hast dadurch plötzlich mehr Deutsch gesprochen? Oder warst du mit anderen internationalen Studenten unterwegs, mit denen du nur noch Englisch gesprochen hast? Wenn sich nichts großartig verändert hat / verändern wird, schreib' einfach "Nein" in das Feld, in allen anderen Fällen gib so viel Infos, wie möglich. [If so, did your language change / will it change? Are you, for example, travelling with German friends, or with international friends, with whom you solely speak English? If nothing changes / changed, just enter 'no', otherwise provide as much information as possible.]

Wie hast du deine Freizeit im Ausland verbracht? Gib im Feld neben jeder ausgewählten Aktivität an, zu wie viel Prozent du deine Freizeit damit gestaltet hast. [How did you spend your freetime abroad? Indicate for each activity to what percent you spent you free time with this activity.]

- Vereinsaktivitäten (Sportclub, Chor, Tanzgruppe, etc.) [social activities in organized groups, such as sports club, choir, dance group etc.]
- Partys / Kneipenabende [partys, bar evenings]
- Reisen [travelling]
- Kulturelle Aktivitäten (Museum, Theater, Konzerte etc.) [cultural activities, such as museum, theatre or concert visits]
- Netflix / TV schauen (allein zuhause) [Netflix, watching TV, alone at home]
- Lesen (Zeitung, Bücher...) [reading book or magazines]
- Anderweitig Zeit mit Freunden (in Situationen die hier nicht aufgelistet sind) [otherwise spent time with friends in situations that are not listed here]

Wann genau hast du Spanien endgültig verlassen, bzw. wirst du Spanien verlassen? [When did you leave Spain / when will you leave Spain for good?]

D.7.3 | *T3-specific Questions*

Beschreibe auf SPANISCH, was deiner Meinung nach die größten Unterschiede zwischen Spaniern und Deutschen sind und warum. Denke hierbei an Mentalität und Lebensweise, aber auch an das Unisystem. Vermisst du die spanische Lebensweise, oder bist du froh wieder in Detuschland zu sein? [Describe in Spanish what you consider to be the biggest differences between Spanish and German people. Think about mentality, way of living, but also the university system. Do you miss the Spanish way of life, or are you happy to be back in Germany?]

In welchen Kontexten verwendest du aktuell noch Spanisch? [In which contexts do you currently still use Spanish?]

Hast du noch Kontakt zu Leuten, die du in Spanien kennen gelernt hast? Wenn ja, mit wem und wie intensiv? [Are you still in touch with people you met in Spain? If so, with whom and how intensively?]

Gab es seit Rückkehr nach Deutschland Phasen mit deutlich mehr (oder deutlich weniger) Spanischkontakt, durch z.B. Besuch aus Spanien, einer Reise nach Spanien oder dergleichen? Beschreibe die Situation kurz und gib eine ungefähre Zeitangabe. [Have there been phases with clearly more (or less) Spanish use since you returned to Germany, for example because of visitors from Spain, or a trip to Spain? Describe the situation briefly and provide an approximate time indication.]

Hast du seit deiner Rückkehr aus Spanien/Lateinamerika noch Unterricht auf Spanisch gehabt (Kurse an der Uni, Sprachkurs)? [Have you followed Spanish courses since you returned from Spain? (e.g., at university, or a language course)]

□ Ja [yes]
□ Nein [no]

Liste hier alle Kurse denen du derzeit (bzw. seit Rückkehr aus Spanien/LA) auf Spanisch folgst / gefolgt hast. [List all courses that you followed in Spanish since returning from Spain.]

Kurstitel [title]	Stunden/Woche [hours/week]	Ungefährer Zeitraum [appr. time range]
...		

Was ist deine aktuelle Wohnsituation? [What is your current living situation?]

□ Eigene Wohnung [own apartment]
□ WG mit Deutschen [shared apartment with Germans]
□ WG mit unter anderen spanischsprachigen Leuten [shared apartment with Spanish-speaking people]
□ WG mit internationalen Leuten (aber nicht spanischsprachig) [shared apartment with non-Spanish speaking internationals]

Inwiefern stimmst du den folgenden Aussagen zu? [To what extent do you agree with the following statements?]

1 (stimme überhaupt nicht zu [absolutely disagree]) – 4 (neutral) – 7 (stimme absolut zu [absolutely agree])

- Ich probiere aktiv mein Spanisch zu erhalten. [I actively try to maintain my Spanish.]
- Ich bin in Deutschland viel mit Spaniern in Kontakt. [In Germany I'm often in touch with Spanish people.]
- Ich schaue Filme/Serien auf Spanisch. [I watch movies and series in Spanish.]
- Ich lese Bücher auf Spanisch. [I read books in Spanish.]

D.7.4 | *Motivation Questionnaire (Identical at T1, T2 and T3)*

Answers were given on a 7-point Likert scale from 1 – stimme überhaupt nicht zu [I do not agree at all] to 7 – stimme vollständig zu [I fully agree]. The table below shows all questions, as well as their category assignment and whether they were negatively phrased or not. The latter two types of information were not provided to participants.

Question	Category	Negative
Ich würde gern viele Fremdsprachen fließend und perfekt sprechen können. [I would like to be able to speak many foreign languages fluently and perfectly.]	interest in foreign languages (FL)	0
Spanier sind sympathisch und gastfreundlich. [Spanish people are likable and welcoming.]	attitude towards Spanish people	0
Spanisch lernen ist wichtig, weil es mir erlaubt mich wohler zu fühlen, wenn ich mit Spaniern spreche. [Learning Spanish is important because it allows me to feel more at ease when I speak to Spanish people.]	Integrative motivation	0
Fremdsprachen lernen ist unangenehm. [Learning foreign languages is uncomfortable.]	interest in FL	1
Spanisch lernen ist wichtig, da ich es für meine berufliche Karriere brauche. [Learning Spanish is important because I need it for my career.]	Instrumental motivation	0
Ich würde gern Zeitungen in vielen verschiedenen Sprachen lesen können. [I would like to be able to read the newspaper in many different languages.]	interest in FL	0
Ich verstehe nicht warum manche Austauschstudenten nervös werden, wenn sie Spanisch sprechen müssen. [I don't understand why some exchange students get nervous when speaking Spanish.]	anxiety	1
Die meisten Spanier sind sympathisch und es ist so einfach sich gut mit ihnen zu verstehen, dass ich froh bin, sie als Freunde zu haben. [Most Spanish people are likable and it is so easy to get along well with them that I'm happy to have them as my friends.]	attitude	0

Question	Category	Negative
Spanisch lernen ist wichtig, da es mir erlaubt mit mehr Leuten in Kontakt zu treten und zu sprechen. [Learning Spanish is important because it enables me to get in touch and speak with more people.]	integrative	0
Ich habe kein Interesse an Fremdsprachen. [I'm not interested in foreign languages.]	interest in FL	1
Ich bin selbstsicher, wenn ich im Restaurant auf Spanisch mein Essen bestelle oder jemandem auf der Strasse den Weg weise. [I'm confident when I order food in Spanish at a restaurant or when I talk to someone on the street to show them the way.]	anxiety	1
Spanisch lernen ist wichtig, da es mich gebildeter macht. [Learning Spanish is important because it makes me more educated.]	instrumental	0
Ich hätte gern viele spanische Freunde. [I would like to have many Spanish friends.]	attitude	0
Ich würde wirklich gern viele Fremdsprachen sprechen können. [I would really like to be able to speak many foreign languages.]	interest in FL	0
Die Spanier sind sehr sozial und liebenswürdig. [The Spanish people are very sociable and loveable.]	attitude	0
Ich habe kein Problem mit Leuten auf der Strasse auf Spanisch zu sprechen. [I have no problem talking to people in Spanish on the street.]	anxiety	1
Spanisch lernen ist wichtig, da es mir dabei helfen wird, die spanische Lebensweise zu leben und zu verstehen. [Learning Spanish is important because it will help me understand and live the Spanish way of life.]	integrative	0
Spanier haben viel worauf sie stolz sein können, da sie viel zur Welt beigetragen haben. [The Spanish people have a lot tob e proud of because they contributed a lot to the world.]	attitude	0
Es ist wichtig Fremdsprachen zu lernen. [It is important to learn foreign languages.]	interest in FL	0
Es ist wichtig Spanisch zu lernen um einen guten Job zu finden. [It is important to learn Spanish to find a good job.]	instrumental	0
Wenn ich längere Zeit in ein fremdes Land gehe, versuche ich die Landessprache zu lernen. [If I move to a foreign country for a longer period of time, I try to learn the language.]	interest in FL	0
Ich würde gern mehr Spanier kennen lernen. [I would like to meet more Spanish people.]	attitude	0
Spanisch lernen ist mir wichtig, weil ich so einfacher mit Spaniern in Kontakt kommen könnte. [Learning Spanish is important to me because it allows me get into touch with Spanish people more easily.]	integrative	0
Viele Fremdsprachen klingen grob und ordinär. [Many foreign languages sound rough and vulgar.]	interest in FL	1
Spanisch lernen ist wichtig, da mich die Leute dann mehr respektieren. [Learning Spanish is important because people will respect me more if I speak it.]	instrumental	0
Ich werde sehr nervös, wenn ich mit Muttersprachlern Spanisch reden muss. [I get nervous when I have to talk to native speakers of Spanish.]	anxiety	0

Question	Category	Negative
Ich lerne gern Leute kennen, die Fremdsprachen sprechen. [I like meeting people who learn foreign languages.]	interest in FL	0
Je mehr ich spanischsprechende Leute kennen lerne, desto besser verstehe ich mich mit ihnen. [The more I get to know Spanish-speaking people, the better I get along with them.]	attitude	0
Ich bevorzuge Filme synchronisiert in meiner eigenen Muttersprache zu gucken, anstatt mit Untertiteln und in Originalsprache. [I prefer watching movies synchronized in my own mother tongue rather than in original version with subtitles.]	interest in FL	1
Man kann sich immer auf Spanier verlassen. [You can always rely on Spanish people.]	attitude	0
Ich fühle mich wohl, wenn ich Spanisch spreche. [I feel comfortable when I speak Spanish.]	anxiety	1
Ich wünschte ich würde perfektes Spanisch sprechen. [I wish I spoke perfect Spanish.]	integrative	0
Questions on individual 7-point scales [see brackets for scales]		
Meine Motivation Spanisch zu lernen um mit Spaniern reden zu können ist: (Niedrig – Hoch) [My motivation to learn Spanish to be able to speak to Spanish people is: low - high]	integrative	0
Mein Verhalten gegenüber Spaniern ist: (unfreundlich – freundlich) [My behavior towards Spanish people is: unfriendly - friendly]	attitude	0
Meine Interesse Fremdsprachen zu lernen ist: (niedrig – hoch) [My interest in learning foreign languages is: low - high]	interest in FL	0
Meine Motivation Spanisch zu lernen aus praktischen Gründen (für einen späteren Job), ist: (schwach – stark) [My motivation to learn Spanish for practical reason (for a job) is: weak - strong]	instrumental	0
Meine Nervosität, wenn ich Spanisch im Alltag spreche, ist: (niedrig – hoch) [My nervosity when I speak Spanish in everyday life is: low - high]	anxiety	0

D.7.5 | *Questions Regarding Language Use: Spanish*

D.7.5.1 | T1-specific

Wie alt warst du, als du angefangen hast Spanisch zu lernen? [How old were you when you started learning Spanish?]

Wie viele Jahre lernst du schon / hast du bisher intensiv Spanisch gelernt? [How many years have you been intensively learning Spanish?]

Mit wem sprichst du Spanisch (in Deutschland)? [Who do you speak Spanish with (in Germany)?]
□ Freunde [friends]
□ Partner [partner]
□ Kommilitonen [fellow students]
□ Mitbewohner [house mates]
□ Ich spreche fast nie Spanisch [I hardly ever speak Spanish]

Wie hast du Spanisch gelernt? [How did you learn Spanish?]
□ In der Schule [at school]
□ An der Uni [at university]
□ Mit Freunden [with friends]
□ Durch einen Auslandsaufenthalt [through a study abroad]
□ Sonstiges [other]

Zu welchem CERF Niveau hat dein letzter Spanischkurs in Deutschland geführt? [Which CERF level did you reach with your latest Spanish language course?]
□ A1
□ A2
□ B1
□ B2
□ C1
□ C2
□ Ich habe bisher keinen Spanischkurs besucht. [I haven't attended any Spanish language courses yet.]

D.7.5.2 | **T2 and T3- specific**

T2: Hast du in Spanien Spanisch gesprochen? [Did you speak Spanish while in Spain?]
□ Ja [yes]
□ Nein [no]

T3: Sprichst du noch regelmäßig Spanisch? [Do you still regularly speak Spanish?] Auch sporadischer Sprachgebrauch ist hier mit gemeint. [Occasional language use also counts.]
□ Ja [yes]
□ Nein [no]

T2 & T3: Zu wie viel Prozent sprichst du aktuell mit den aufgeführten Personengruppen Spanisch? Die Summe sollte 100% ergeben. 100% reflektiert die Gesamtheit deiner Zeit, die du Spanisch sprichst, also auch wenn das stundenmäßig wenig war, geht es uns darum, wie viel von dieser Zeit du mit wem sprichst. [What percent of the time do you currently speak Spanish with these two groups of people? The numbers should add up to 100%. 100% reflects the entire time that you speak Spanish. Even if you speak little Spanish, we want to know to what extent you do so with whom.]

□ Spanische Muttersprachler [Spanish native speakers]
□ Nicht-Muttersprachler [non-native speakers of Spanish]

T2: Hast du in Spanien einen Spanischkurs besucht? [Did you follow a Spanish language course while abroad?]

Zu welchem CERF Level hat dieser Kurs geführt? Und wie viele Stunden pro Woche dauerte er? [Which CERF level did this course lead to and how many hours a week did it take?]

D.7.5.3 | **Identical for T1, T2 and T3**

Wie schätzt du aktuell deine Spanischkenntnisse auf den folgenden Gebieten ein? [How do you currently judge your Spanish proficiency in these four domains?]

1 (sehr schlecht [very poor]) – 7 (sehr gut, auf Muttersprachlerniveau [very good, on a native level])

- Lesen [reading]
- Schreiben [writing]
- Hören [listening]
- Sprechen [speaking]

T2 & T3 only: Würdest du sagen, dass sich dein Spanisch [T2: während deines Auslandsaufenthaltes, T3: seit Rückkehr aus Spanien/Lateinamerika] verbessert oder verschlechtert hat? [Would you say that your Spanish improved or got worse [T2: during your study abroad, T3: since you returned from Spain]?]

1 – stark verschlechtert [got a lot worse] – 4 keine Veränderung [no change] – 7 stark verbessert [got a lot better]

D.7.6 | *Questions regarding language use: English*

D.7.6.1 | **T1-specific**

Wie alt warst du, als du angefangen hast Englisch zu lernen? [How old were you when you started learning English?]

Wie viele Jahre lernst du schon / hast du bisher intensiv Englisch gelernt? [How many years have you been intensively learning English?]

Mit wem sprichst du Englisch (in Deutschland)? [Who do you speak English with (in Germany)?]
□ Freunde [friends]
□ Partner [partner]
□ Kommilitonen [fellow students]
□ Mitbewohner [house mates]
□ Ich spreche fast nie Englisch [I hardly ever speak English]

Wie hast du Englisch gelernt? [How did you learn English?]
□ In der Schule [at school]
□ An der Uni [at university]
□ Mit Freunden [with friends]
□ Durch einen Auslandsaufenthalt [through a study abroad]

D.7.6.2 | **T2 & T3-specific**

T2: Hast du in Spanien Englisch gesprochen? [Did you speak English in Spain?]
□ Ja [yes]
□ Nein [no]

T3: Sprichst du noch regelmäßig Englisch? [Do you still regularly speak English?] Auch sporadischer Sprachgebrauch ist hier mit gemeint. [Occasional language use also counts.]
□ Ja [yes]
□ Nein [no]

T2 & T3: Zu wie viel Prozent sprichst du aktuell mit den aufgeführten Personengruppen Englisch? Die Summe sollte 100% ergeben. 100% reflektiert die Gesamtheit deiner Zeit, die du Englisch sprichst, also auch wenn das stundenmäßig wenig war, geht es uns darum, wie viel von dieser Zeit du mit wem sprichst. [How much percent of the time do you currently speak English with these two groups of people? The numbers should add up to 100%. 100% reflects the entire time that you speak English. Even if you speak little English, we want to know to what extent you do so with whom.]
□ Englische Muttersprachler (English native speakers)
□ Nicht-Muttersprachler (non-native speakers of English)

D.7.6.3 | **Identical for T1, T2 and T3**

Wie schätzt du aktuell deine Englischkenntnisse auf den folgenden Gebieten ein? [How do you currently judge your English proficiency in these four domains?]

1 (sehr schlecht [very poor]) – 7 (sehr gut, auf Muttersprachlerniveau [very good, on a native level])

- Lesen [reading]
- Schreiben [writing]
- Hören [listening]
- Sprechen [speaking]

D.7.6.4 | **T2 & T3 only**

Würdest du sagen, dass sich dein Englisch [T2: während deines Auslandsaufenthaltes, T3: seit Rückkehr aus Spanien/Lateinamerika] verbessert oder verschlechtert hat? [Would you say that your English improved or got worse [T2: during your study abroad, T3: since you returned from Spain]?]

1 – stark verschlechtert [got a lot worse] – 4 keine Veränderung [no change] – 7 stark verbessert [got a lot better]

D.7.7 | *Questions Regarding Language Use: Other Foreign Languages*

Wie schätzt du aktuell deine Sprachkentnisse in den folgenden Sprachen ein? [How do you currently judge your proficiency in the following languages?]

1 sehr schlecht (very poor) – 7 sehr gut, auf Muttersprachlerniveau (very good, native-like)

- Französisch [French]
- Italienisch [Italian]
- Portugiesisch [Portuguese]
- Latein [Latin]
- Russisch [Russian]
- Katalanisch [Catalan]

D.7.8 | *Frequency of use questions (identical at T1, T2, and T3)*

Gib an zu wie viel Prozent du aktuell die folgenden Sprachen in den angegebenen Kontexten verwendest. Jede Reihe sollte insgesamt 100% ergeben. [Indicate what percent of your time you currently spend using the following languages in the following contexts. The sum of each row should add up to 100%.]

	Spanisch [Spanish]	Englisch [English]	Deutsch [German]	Andere [other]
SPRECHEN (z.B. mit Freunden, Familie, an der Uni, im Pub, im Restaurant, beim Einkaufen…) [SPEAKING (e.g., with friends, family, at university, in the pub, at the restaurant, while shopping…)]				
SCHREIBEN (z.B. E-Mails, Briefe, Formulare, Prüfungen, im Internet…) [WRITING (e.g., email, letters, exams, online…)]				
HÖREN (z.B. an der Uni, im Alltag, im Radio, Fernsehen, mit Freunden oder Familie …) [LISTENING (e.g., at university, on the radio, tv, with friends and family …)]				
LESEN (z.B. im Internet, Bücher, E-Mails…) [READING (e.g., online, books, emails…)]				

Falls du bei der obigen Frage auch den Gebrauch anderer Sprachen angegeben hast, welche anderen Sprachen sind das? [If you indicated the use of other languages above, which languages are those?]

D.7.9 | *Language Use in Hours (only at T2 and T3)*

Gib an wie viele Stunden du die folgenden Sprachen in den angegeben Kontexten an einem durschnittlichen Tag verwendest. Gehe von der letzten Woche aus. Du kannst auch Bruchteile einer Stunde angeben, also 0.5 für 30 Minuten. [Indicate how many hours you use the following languages in the following contexts on an average day. Use the last week as a reference. You can also indicate fractions of an hours, e.g., 0.5 for 30 minutes.]

	Spanisch [Spanish]	Englisch [English]	Deutsch [German]	Andere [other]
Lesen von Büchern, Zeitschriften, Webseitinhalten etc. [Reading of books, magazines, website content etc.]				
Filme schauen, Radio/ Podcasts / Hörbücher hören [watching movies, listening to radio / podcasts / audio books]				

Kurse an der Uni / Interaktion mit anderen Studenten an der Uni [Courses at university, interaction with other students at uni]				
Interaktion mit Freunden (auf Partys, in Bars, im Café, Dates etc.) [Interaction with friends, at parties, in bars, at cafes, on dates, etc.]				
Interaktion mit Mitbewohnern [interaction with house mates]				
Interaktion mit Partner(in) [interaction with partner]				
T2-specific: Interaktion mit Freunden und Familie in Deutschland [interaction with friends and family in Germany]				
T3-specific: Interaktion mit Freunden in Spanien [interaction with friends in Spain]				

Hast du noch Kommentare zu deinem Sprachgebrauch? [Do you have any other comments on your language use?]

D.7.10 | *Language Use by Groups of People (only at T2 and T3).*

Welche Sprachen hast du mit den aufgelisteten Personengruppen [T2: im Ausland; T3: in Deutschland] gesprochen? Gib pro Gruppe an, zu wie viel Prozent du mit dieser Gruppe die entsprechenden Sprachen gesprochen hast. Pro Gruppe sollte die Summe 100 % betragen. Die Summe pro Sprachspalte kann mehr als 100% sein. Wenn du beispielsweise mit Spaniern immer Spanisch gesprochen hast, und mit internationalen Studenten 20% der Zeit Spanisch gesprochen hast, wäre die Summe der Spanischspalte 120 - das ist völlig ok, solange die Summe pro Reihe, also pro Gruppe 100% ergibt. [Which languages did you use with which group of people [T2: while abroad, T3: while back in Germany]? For each group, indicate what percentage of time you spoke to this group in each of the languages. The sum of percentages per group should be 100%. The sum of all percentages for the different languages can in turn be higher than 100%. If you, for example, always spoke Spanish with Spanish people and with international people you spoke Spanish 20% of the time, the sum for Spanish would be 120 %. This is ok, as long as the sum for each group equals 100%.]

	Spanisch [Spanish]	Englisch [English]	Deutsch [German]	Andere [other]
Spanische Muttersprachler [Spanish native speakers]				
Deutsche Muttersprachler [German native speakers]				
Andere (nicht Deutsche) international Studenten [other, non-German international students]				

D.8 | Correlation Matrix for Frequency of Use Subscores

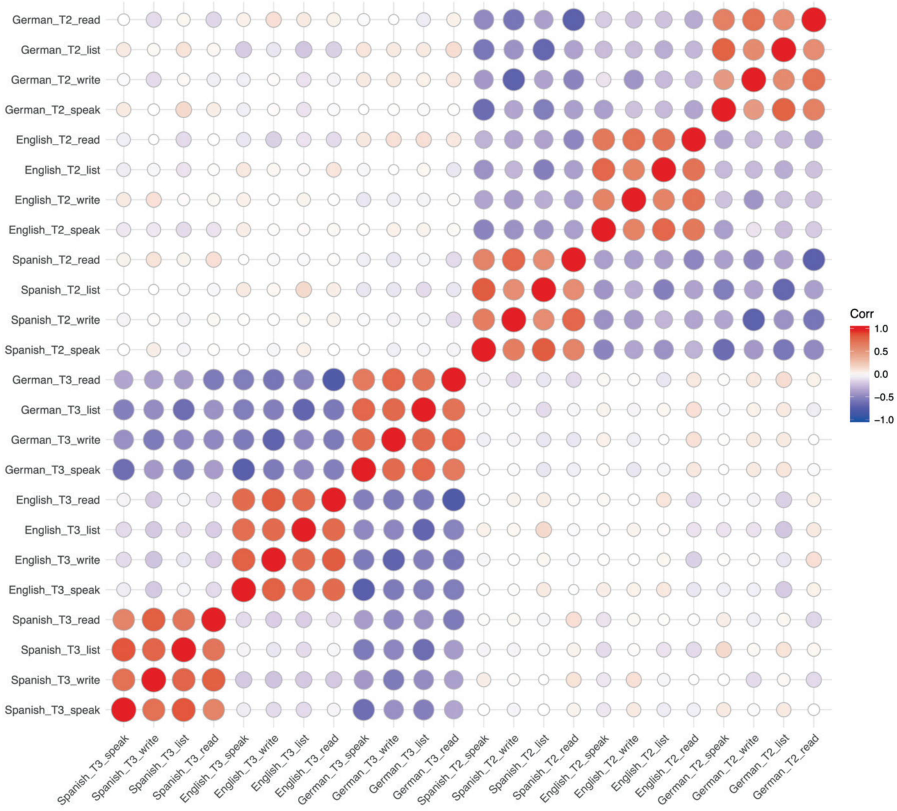

Figure D.8

Pearson correlation matrix for frequency of use measures per domain at T2 and T3. Color and circle size reflect the strength of the correlation with shades of red indicating positive and shades of blue indicating negative correlations.

D.9 | Correlation Matrix for Motivation Questionnaire Subscores

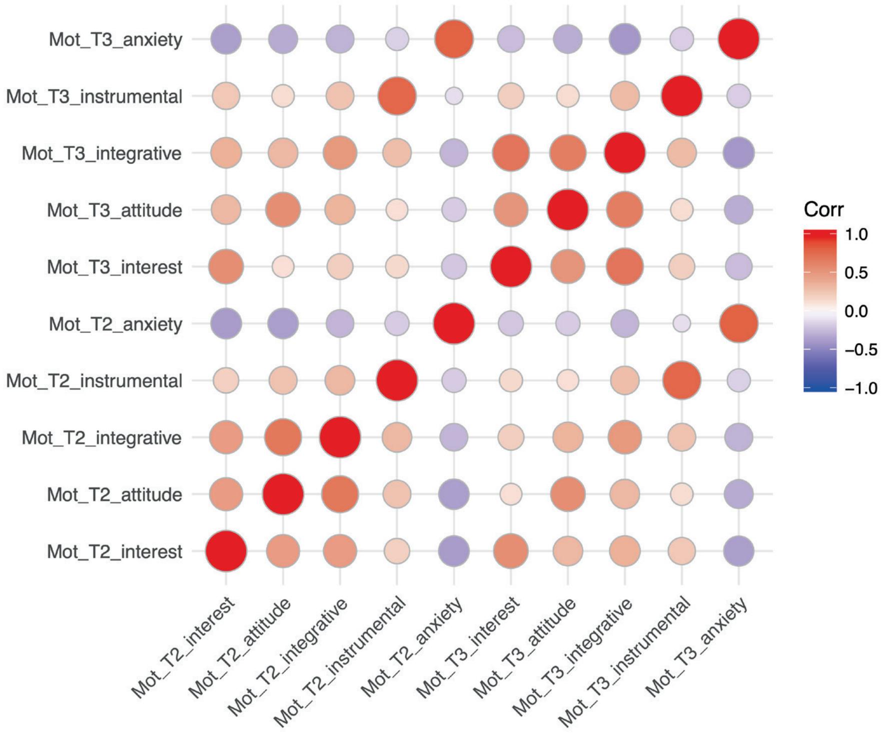

Figure D.9

Pearson correlation matrix for motivation questionnaire scores per question category at T2 and T3. Color and circle size reflect the strength of the correlation with shades of red indicating positive and shades of blue indicating negative correlations.

D.10 | Doors Stimuli

From the available set of 100 door display pairs (target displays and 4 AFC displays), we selected the following 30 pairs (numbers refer to slide numbers in the original PowerPoint files, available here: https://www.york.ac.uk/res/doors/resources.shtml): 1, 2, 9, 10, 11, 12, 19, 21, 21, 22, 87, 88, 27, 28, 41, 42, 47, 48, 51, 52, 53, 54, 71, 75, 76, 81, 82, 85, 86, 91, 92, 97, 98, 99, 100, 123, 124, 129, 130, 133, 134, 141, 142, 143, 144, 145, 146, 155, 156, 169, 170, 179, 180, 181, 182, 193, 194, 195, 196.

D.11 | Correlation Matrix for Predictors Including Input Type

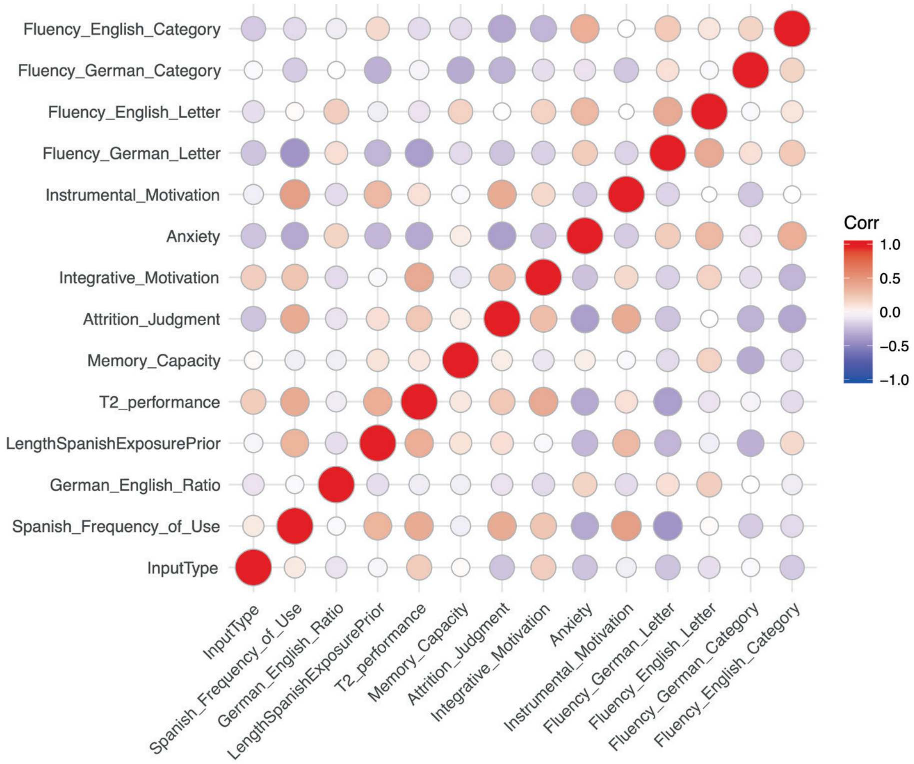

Figure D.11

Pearson correlation matrix for all 14 participant-level predictors from the additional analysis with Spanish input type (% native) as extra predictor (N = 49). Colors indicate the strength of the correlation (Pearson's *r*) with shades of blue indicating negative and shades of red indicating positive correlations.

D.12 | Predictor Plots for Relative Forgetting Rates

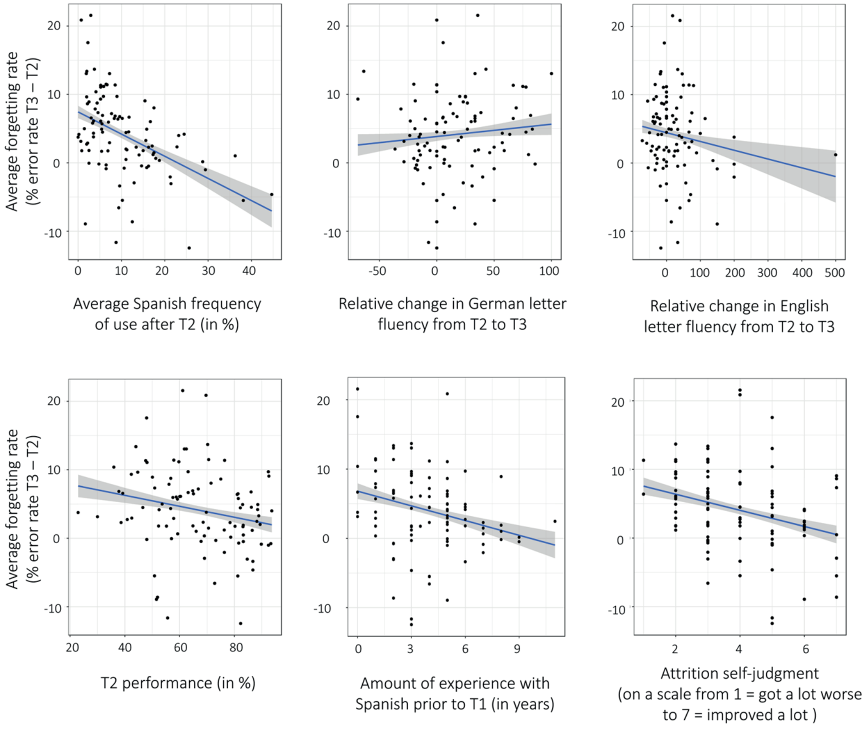

Figure D.12

Participant-level predictor plots. The y-axis plots the average difference in error rates from T2 to T3 for each participant, such that a positive difference reflects forgetting from T2 to T3, while a negative difference reflects learning from T2 to T3. Dots represent individual participants. The x-axes plot the respective participant-level predictors. Lines reflect the best-fit linear relationship between each predictor and the T2-T3 change in error rates with 95% confidence intervals. This was done to provide a simpler visualization of the significant Session*Predictor interactions in the GLMM reported in the main text. Note that the effect of T2 performance on forgetting rates is different in this plot than in the GLMM reported in the main text, because here we plot forgetting rates relative to each participant's baseline T2 performance, while the GLMM evaluates the effect that T2 performance has on absolute differences in performance between T2 and T3, which turns out to be opposite (see section 5.4.2 for more details). Given that T2 performance is included as a predictor in the main model though, the model implicitly takes T2 performance into account in evaluating the effects of all other predictors in the model. For those other predictors, this plot thus shows the same type of relationship to forgetting rates as the GLMM in the main text.

Nederlandse Samenvatting

Laatst kreeg ik een onverwacht telefoontje van een oude vriendin. Cristina en ik hadden elkaar tijdens mijn bachelor studie in Berlijn in 2009 leren kennen. Ik was net een studie Spaans begonnen en zij, oorspronkelijk uit Spanje, was op Erasmusuitwisseling. Tegen het einde van onze gezamenlijke tijd in Berlijn had ik er niet alleen een hele goede vriendin bij, maar ik had ook zoveel tijd met haar en haar Spaanse vrienden doorgebracht dat ik bijna vloeiend Spaans sprak. Tegenwoordig spreek ik helaas bijna nooit meer Spaans, en toen Cristina mij onlangs belde werd mijn gebrek aan oefening op pijnlijke wijze duidelijk. Ik was constant op zoek naar woorden en uitdrukkingen in het Spaans. Om toch te praten schakelde ik continu over naar het Engels, en tot Cristina's verwarring, glipten er ook regelmatig Nederlandse woorden in mijn zinnen zonder dat ik het zelf merkte.

Het was natuurlijk al een tijdje geleden dat ik voor het laatst Spaans had gesproken, toch was het verbijsterend om te zien hoeveel moeite ik had een taal te spreken die ik ooit zo goed had beheerst. Natuurlijk ben ik niet de enige met deze ervaring. De meeste mensen die een vreemde taal hebben geleerd en hem vervolgens niet meer gebruiken, zullen dit vervelende gevoel kennen. Hoe komt het dat we vreemde talen zo makkelijk vergeten, en wat bepaalt hoe snel en hoeveel we van de taal verliezen? Gaan taalvaardigheden achteruit omdat we een taal lang niet gebruiken, of zijn er nog andere processen bij betrokken? In dit proefschrift heb ik antwoorden op deze vragen gezocht.

Om het vergeten van vreemde talen te studeren, heb ik me laten inspireren door onderzoek uit de algemene geheugenliteratuur. We vergeten namelijk niet alleen talen, maar ook andere dingen, zoals: waar we ons fiets hebben geparkeerd, wat we behalve melk en eieren verder nog in de supermarkt wilden kopen of hoe onze basisschool-klasgenoot die zo ontzettend goed kon tekenen heette. Onderzoek naar hoe vergeten werkt gaat terug naar de 19e eeuw. Onderzoekers hebben sindsdien een aantal theorieën bedacht over waarom mensen überhaupt vergeten. De meest bekende en wetenschappelijk ondersteunde theorie is de interferentie theorie. In plaats van 'vergeten' te verklaren als het bijproduct van de voortgang van tijd, stelt de interferentie theorie dat 'vergeten' tot stand komt door competitie en dus door interferentie van andere herinneringen. Informatie wordt nooit geïsoleerd in ons geheugen opgeslagen maar altijd in een groot netwerk, waar herinneringen die op elkaar lijken met elkaar verbonden zijn. Deze verbindingen leiden tot co-activatie: als je de naam van een oude klasgenoot herinnert zullen ook de namen van andere klasgenoten omhoogkomen. De gerelateerde herinneringen staan dan vaak in de weg en moeten onderdrukt worden om de juiste informatie te kunnen selecteren. Terwijl het onderdrukken van onnodige informatie bij het herinneren helpt, blijkt uit onderzoek in de geheugenliteratuur dat het tegelijkertijd ten koste gaat van de latere toegankelijkheid van deze onderdrukte informatie. De onderdrukte informatie wordt moelijker op te halen. Hoe vaker een bepaalde herinnering in de weg staat en vervolgens onderdrukt wordt, hoe lastiger het wordt deze later op te halen. Op een gegeven moment lukt het helemaal niet meer en lijkt de informatie dus vergeten. In de geheugenliteratuur wordt er echter niet van uitgegaan dat informatie helemaal verdwijnt, maar dat vergeten informatie alleen tijdelijk niet bereikbaar is.

Uit de algemene geheugenliteratuur blijkt dus dat herinneren tot vergeten kan leiden, namelijk tot het vergeten van informatie die gerelateerd is aan de herinnerde informatie, een fenomeen dat ook 'retrieval-induced forgetting' wordt genoemd. Er bestaat ook nog 'retroactieve interferentie', die ontstaat als je nieuwe informatie toevoegt. Bijvoorbeeld: als je je wilt herinneren wie in klas 5 naast je zat op school is dat wellicht moeilijk omdat je in de navolgende jaren naast iemand anders zat en je het concept 'klasgenoot' vooral met deze persoon associeert in plaats van met die persoon uit klas 5. Welke soort interferentie je ook ervaart, je vergeet dus niet alleen omdat je informatie al lang niet naar boven hebt gehaald, maar ook -en misschien vooral- omdat je in de tussentijd andere, gerelateerde informatie hebt gebruikt of geleerd.

In mijn promotieonderzoek heb ik onderzocht of interferentie ook tussen talen bestaat en of het wellicht één van de mechanismen achter het vergeten van taal is. Onderzoek met meertaligen heeft al eerder aangetoond dat alle talen die je spreekt tot op zekere hoogte tegelijk actief zijn. Stel dat Nederlands je moedertaal is en je ook Engels en Spaans spreekt. Als je dan met je Spaanse vriendin over je laatste fietstocht praat, is niet alleen het Spaanse woord 'bicicleta' actief, maar ook het Engelse woord 'bicycle' en natuurlijk het Nederlandse woord fiets. Woorden uit andere talen worden kennelijk automatisch en onbewust mede geactiveerd, omdat ze naar hetzelfde onderwerp verwijzen. Net als de herinneringen over je klasgenoten zijn vertalingen dus aan elkaar gerelateerd. Ze kunnen daardoor wellicht met elkaar interfereren. Als dit het geval is, zou het kunnen zijn dat we vreemde talen vergeten omdat we andere talen in de tussentijd gebruiken of leren. In hoofdstukken 2 tot en met 5 testte ik op verschillende manieren of dit het geval is.

Het idee dat interferentie één van de oorzaken is van het vergeten van taal, is niet helemaal nieuw. We weten uit observaties dat, naarmate de tijd verstrijkt, mensen steeds vaker woorden uit de wegslijtende taal vervangen door woorden uit de taal van hun actuele omgeving. Op basis van deze observaties hebben onderzoekers al lang geleden voorgesteld dat het vergeten van een taal door het gebruik van andere talen veroorzaakt wordt. Dat je woorden in een vreemde taal vervangt door anderstalige woorden toont echter niet aan dat interferentie tussen talen de oorzaak van vergeten is. Het vervangen van woorden is alleen het gevolg van het vergeten. In onderzoek waar je iemands taalvaardigheid op verschillende momenten toetst kun je vaststellen of er met de tijd wel of niet sprake van verlies is, maar je hebt geen toegang tot de tijd tussen de toetsen, waar het vergeten eigenlijk gebeurt. Omdat traditioneel onderzoek naar het vergeten van taal dus alleen de uitkomst maar niet het vergeetproces zelf kan bestuderen, weten we nog steeds heel weinig over wat precies tot het vergeten van vreemde talen leidt. In mijn onderzoek gebruikte ik daarom een andere methode, geïnspireerd door onderzoek in de algemene geheugenliteratuur. In plaats van het vergeten van taal in het echte leven te observeren, heb ik in **hoofdstukken 2 tot en met 4** geprobeerd het vergeten van talen in het lab te simuleren. Ik heb onderzocht welke interventies bij proefpersonen leiden tot het vergeten van woorden uit een vreemde taal. Door te kijken welke interventies wel en welke niet tot (tijdelijk) vergeten leidden, kon ik uiteindelijk conclusies trekken over de mechanismen die tot het vergeten van een vreemde taal kunnen leiden.

In **hoofdstuk 2** liet ik mensen eerst door middel van verschillende taken een aantal nieuwe woorden leren in een vreemde taal (Spaans, L3). Het doel was dat mijn Nederlandstalige proefpersonen aan het einde van deze leersessie een aantal plaatjes van alledaagse voorwerpen, bijvoorbeeld het plaatje van een lepel of een riem, makkelijk en zonder fouten in het Spaans konden benoemen. Vervolgens probeerde ik ze deze woorden te laten vergeten door ze de helft van deze plaatjes of in hun moedertaal Nederlands of in het Engels te laten benoemen. Ik deelde mijn proefpersonen dus in twee groepen: de ene groep kreeg interferentie in het Nederlands, de andere in het Engels. Als interferentie van het recente gebruik van andere talen tot vergeten zou leiden, dan zou je verwachten dat mijn proefpersonen in een latere test meer moeite hebben zich de Spaanse woorden te herinneren die ze recent in het Engels of Nederlands hebben moeten benoemen dan de woorden die ze tussentijds niet moesten benoemen.

In **hoofdstuk 2** liet ik zien dat dit inderdaad het geval was: deelnemers waren langzamer en minder nauwkeurig in het benoemen van Spaanse woorden die ze tussendoor in het Nederlands of in het Engels hadden benoemd. Dit interferentie-effect hield aan tot een week na de interferentie. Bovendien bleek het interferentie-effect iets sterker voor de Engelse dan voor de Nederlandse interferentiegroep: mensen die tussendoor plaatjes in het Engels moesten benoemen hadden later meer moeite met het zich herinneren van de Spaanse woorden dan mensen die de plaatjes in hun moedertaal Nederlands moesten benoemen. Het tussentijds gebruiken van een andere vreemde taal blijkt dus iets nadeliger te zijn voor het latere L3-herinneringsvermogen dan interferentie van het gebruik van je moedertaal. Het tweede experiment in **hoofdstuk 2**, suggereert dat dit waarschijnlijk te wijten is aan verschillen in gebruiksfrequentie tussen moedertaal en niet-moedertaal. Woorden uit minder vaak gebruikte talen (zoals Engels voor mijn proefpersonen) zijn moeilijker op te halen en eisen daardoor meer inhibitie van storende vertalingen (zoals Spaanse woorden in mijn experiment). Minder gebruikte talen blijken daardoor sterker te interfereren dan vaak gebruikte talen. Samengevat toont **hoofdstuk 2** dus aan dat interferentie door het recente gebruik van andere talen inderdaad tot het vergeten van dezelfde woorden in een vreemde taal kan leiden. In tegenstelling tot traditioneel onderzoek naar het vergeten van taal tonen deze twee experimenten een causaal verband tussen taal interferentie en het vergeten van taal aan.

In **hoofdstuk 3** heb ik de resultaten uit **hoofdstuk 2** gerepliceerd, dit keer met Italiaans als nieuw-geleerde taal en Engels als enige interferentie-taal: Nederlandse deelnemers waren ook hier langzamer en minder nauwkeurig in het herinneren van Italiaanse woorden als ze deze recent in het Engels hadden benoemd. Daarnaast keek ik in **hoofdstuk 3** wat er in de hersenen gebeurt wanneer we woorden proberen te herinneren. Kunnen we interferentie tussen talen terugvinden en meten in de breinactiviteit? Om deze vraag te beantwoorden maakte ik gebruik van het feit dat hersencellen continu elektrische signalen uitzenden. De elektrische activiteit van onze hersenen kunnen we door middel van kleine electroden op de schedel opmeten en vervolgens visualiseren en analyseren. Het elektrische signaal wat deze electroden opnemen wordt elektroencephalogram (EEG) genoemd. Op basis van de EEG-metingen, die ik opnam terwijl mijn proefpersonen plaatjes in het Italiaans aan het benoemen waren, constateerde ik dat interferentie tussen talen gekenmerkt wordt door twee patronen: 1. een vroege negatieve piek in het gemiddelde EEG-signaal, ook N2 genoemd, en 2. een toename van activiteit in het theta frequentieband van hersengolven (4-7 Hz). Beide patronen zie je ook in andere (niet taal-gerelateerde) taken terug wanneer er sprake van interferentie, competitie en inhibitie is. Dat mijn proefpersonen een sterkere N2 en sterkere theta band oscillaties lieten zien bij het ophalen van geïnterfereerde Italiaanse woorden in vergelijking met niet geïnterfereerde Italiaanse woorden, komt overeen met het idee dat ze bij het opzoeken van woorden in het Italiaans competitie van de recent gebruikte Engelse woorden ervoeren. Een derde patroon in het EEG was een latere positieve piek (LPC genoemd) die voor de geïnterfereerde woorden minder sterk was dan voor de niet geïnterfereerde woorden. Op basis van de studies die dit patroon eerder hadden gevonden, blijkt het dat de LPC in onze studie de gevolgen van interferentie weergeeft, namelijk: een verminderde toegang tot de geïnterfereerde Italiaanse woorden ten opzichte van de niet geïnterfereerde Italiaanse woorden. Bovendien was ik in staat om activiteit tijdens de interferentiefase in het Engels te koppelen aan de terugvindsnelheid tijdens de eindtoets in het Italiaans. Italiaanse woorden waarvoor de deelnemers langer nodig hadden om ze terug te vinden, en die dus sterker waren verstoord, vertoonden tijdens de eerdere toets in het Engels een hogere N2-piek (meer indicatie voor interferentie en inhibitie) dan de Italiaanse woorden die later gemakkelijker toegankelijk waren. De moeite met het terughalen van woorden ontstaat dus niet pas gedurende de eindtest, maar wordt – zoals interferentietheorie stelt – al tijdens de voorafgaande interferentiefase in gang gezet.

Samen laten **hoofdstukken 2 en 3** zien dat het gebruiken van woorden uit andere talen tot het vergeten van diezelfde woorden in een recent geleerde vreemde taal kan leiden. In **hoofdstuk 4**, laat ik zien dat het tegendeel daarvan ook waar is: het leren van nieuwe Spaanse woorden kan het ophalen van dezelfde woorden in een al lang bekende vreemde taal (Engels voor mijn Nederlandse deelnemers) belemmeren. De negatieve na-effecten van het leren van Spaans waren zichtbaar in zowel de snelheid van het benoemen van plaatjes in het Engels als ook in het herinneringsvermogen van mijn proefpersonen, en ze waren direct na het leren te zien en werden niet sterker met de tijd: het benoemen van plaatjes in het Engels was even moeilijk voor mensen die de toets direct na de Spaanse leersessie moesten doen als voor mensen die de eindtoets pas een dag later hadden. Deze bevinding suggereert dat nieuw geleerde woorden niet eerst in het lange-termijn geheugen hoeven worden opgeslagen om met andere talen te kunnen interageren. Er moet echter nog wel meer onderzoek gedaan worden om dit met zekerheid te kunnen zeggen.

Op basis van de studies in **hoofdstukken 2 tot en met 4** kan ik concluderen dat het vergeten van vreemde talen in het lab geïnduceerd kan worden *door het recente gebruik* van dezelfde woorden in andere talen of *door het leren* van dezelfde woorden in een nieuwe taal. In plaats van alleen te laten zien dat er een correlatie bestaat tussen taalgebruik en het vergeten van een vreemde taal, heb ik met deze studies laten zien dat er inderdaad een *causaal* verband bestaat tussen taalgebruik en taalverlies. Door dit aan te tonen kunnen we nu met meer zekerheid zeggen dat interferentie tussen talen bestaat en op den duur bijdraagt aan het vergeten van taal.

Natuurlijk zijn de lab studies in **hoofdstukken 2, 3 en 4** redelijk artificieel en zou je je kunnen afvragen in hoeverre ze de werkelijkheid modeleren. Het leren van de vreemde taal is bijvoorbeeld sterk vereenvoudigd: het gaat alleen om woorden die mijn proefpersonen binnen een uur of twee hebben geleerd via foto's van de voorwerpen. Verder heb ik bewust geprobeerd om verschillen tussen proefpersonen, bijvoorbeeld in motivatie of eerdere kennis van vreemde talen, uit te schakelen. Dat was nodig om het effect van interferentie in isolatie te onderzoeken, maar deze reductie gaat wel ten koste van de ecologische validiteit. Om het vergeten van taal ook op een natuurlijkere manier te studeren heb ik in **hoofdstuk 5** voor de traditionele aanpak gekozen en volgde ik een groep van Duitse studenten tijdens en na een studieverblijf in Spanje. Ik keek hoe hun taalvaardigheid in het Spaans – met name hun vocabulairekennis - in de eerste zes maanden na hun studieverblijf, toen ze terug in Duitsland waren, veranderde. Ik ging ervan uit dat niet iedereen evenveel Spaans zou onthouden of vergeten, en hoopte dus verschillen tussen mijn proefpersonen te zien in hun vocabulaireverlies. Het hoofddoel was om te onderzoeken welke factoren invloed hebben op hoeveel Spaans iemand binnen zes maanden kwijtraakt. Op basis van mijn lab studies verwachtte ik dat hun Spaanse, Duitse en Engelse taalgebruik één van de bepalende factoren zou zijn. Maar ik keek ook naar motivatie, houding tegenover de Spaanse cultuur, aantal jaren ervaring met Spaans voor het studieverblijf in het buitenland, het bereikte taalniveau aan het einde van de studietijd in Spanje en hun algemene geheugencapaciteit. Leeftijd, socio-economische achtergrond en taalleer-omstandigheden zijn ook mogelijke factoren, maar die heb ik in mijn populatie zo constant mogelijk gehouden en dus niet verder onderzocht.

Op basis van mijn studies in **hoofdstukken 2 – 4** verwachtte ik dat mensen die terug in Duitsland nog regelmatig Spaans gebruiken minder verlies zouden ervaren dan mensen die ineens veel minder Spaans spreken en horen. Omgekeerd betekent dat (zoals mijn lab studies hebben laten zien) dat iemand die vaak andere talen dan Spaans spreekt meer zou moeten verliezen dan iemand die dat minder doet (en dus in verhouding meer Spaans spreekt). Terwijl dit heel logisch lijkt hebben eerdere studies dit verband vaak niet kunnen aantonen. Zoals ik in **hoofdstuk 5** uitleg ligt dat wellicht aan hoe eerdere studies taalgebruik hebben gemeten: slechts één keer, nadat het vergeten van taal al was gebeurd en op abstracte scala's (van 'heel vaak' tot 'heel weinig'). Ik nam in plaats daarvan meerdere, maandelijkse vragenlijsten met antwoorden in procenten af (bijvoorbeeld: op dit moment gebruik ik op een gemiddelde dag 10% Spaans, 70% Duits en 20% Engels). Op basis van deze metingen kon ik laten zien dat ook in natuurlijke situaties een sterke samenhang tussen taalgebruik en het onderhoud van taalvaardigheden bestaat. Duitse studenten die niet meer regelmatig Spaans spraken toen ze weer thuis waren vergaten meer dan Duitse studenten die nog steeds af en toe Spaans gebruikten. Het maakte niet uit of iemand meer Duits of meer Engels praatte: mensen die voornamelijk Duits praatten toen ze terug in Duitsland waren (het meervoud van mijn proefpersonen) vergaten net zo veel als mensen die evenveel Duits en Engels praatten. Er was dus geen bewijs dat vooral het gebruik van andere vreemde talen (Engels) tot vergeten leidt. Omdat geen van mijn proefpersonen meer Engels dan Duits praatte hebben we echter niet genoeg verschillende datapunten om dit met zekerheid te kunnen zeggen. Toekomstig onderzoek zal dit verder moeten bekijken, wellicht in een groep van Duitse studenten die na afloop van hun tijd in Spanje naar Engeland verhuizen en daardoor meer Engels dan Duits praten – een vergelijk tussen de groep uit onze dataset en zo'n Engels-dominante groep zou een betere test van de lab resultaten in **hoofdstuk 2** zijn.

Verder zou je op basis van mijn lab-studies verwachten dat een achteruitgang in Spaanse taalvaardigheid zou samengaan met een verbetering van de verbale vaardigheid in ándere talen. Om dit na te gaan liet ik mensen spreekvaardigheidstests in het Duits en Engels doen: gedurende één minuut moesten mijn proefpersonen zo veel mogelijk woorden die met een bepaalde letter beginnen of deel uitmaken van een bepaalde categorie (bijvoorbeeld 'fruit') opnoemen. Hoe meer woorden je kunt opnoemen, hoe vloeiender je spreekvaardigheid is volgens deze test. Mijn proefpersonen voerden deze taken uit zowel aan het einde van hun studieverblijf in Spanje alsook zes maanden later, terug in Duitsland. Uit mijn resultaten bleek dat mensen bij wie na verloop van tijd een verbetering waarneembaar was in hun Duitse spreekvaardigheid, het sterkste verlies in hun Spaanse vocabulaire kennis vertoonden. Dit komt overeen met mijn lab-resultaten. Voor het Engels was dit niet zo: mensen die in het Engels beter bleken te worden, vertoonden juist minder verlies in hun Spaanse vocabulaire kennis. Dit staat in conflict met de resultaten van mijn lab-studies. Zoals ik in **hoofdstuk 5** uitgebreid bediscussieer, kan het zijn dat de verbale vaardigheidstesten niet de beste meting van taalvaardigheid zijn en daarom niet dezelfde relatie lieten zien. Toen we naar de zelf geschatte taalvaardigheid in het Engels keken zagen we wel een patroon dat op onze lab-studies leek: hoe hoger het (geschatte) verlies in Engelse taalvaardigheid hoe kleiner het (geobserveerde) verlies in de Spaanse vocabulaire kennis.

Naast de kwantiteit was de *kwaliteit* van taalgebruik ook heel erg belangrijk: onafhankelijk van hoeveel Spaans iemand sprak, verloren mensen die vooral input kregen van moedertaalsprekers van het Spaans minder Spaanse woorden - waarschijnlijk omdat zij betere en meer betrouwbare input hadden dan mensen die vooral met andere leerlingen praatten. Ten slotte lieten de data ook nog zien dat meer ervaring met een vreemde taal voordat onze proefpersonen naar het buitenland gingen en een hogere prestatie aan het einde van de tijd in het buitenland ook een positief invloed op het onderhouden van een taal hadden. Daarentegen bleken algemene geheugencapaciteit, motivatie en houding tegenover de Spaanse cultuur geen invloed te hebben op veranderingen in de Spaanse taalvaardigheid van mijn proefpersonen.

Het verlies van vreemde taalvaardigheid is een frustrerend maar veelvoorkomend verschijnsel bij meertaligen. In dit proefschrift onderzocht ik waarom we woorden uit vreemde talen vergeten. Ik benaderde deze vraag vanuit twee invalshoeken: enerzijds door gecontroleerde experimenten die erop gericht waren het vergeten van taal in het lab te simuleren, en anderzijds door in een longitudinale observatie studie studenten te volgen die op natuurlijke wijze hun taalkennis verloren. Samen tonen mijn studies aan dat taalgebruik en competitie tussen talen in ieder geval één van de oorzaken van het vergeten van taal is. Zowel het recente gebruik van al lang gekende talen als ook het leren van een nieuwe taal kunnen leiden tot het vergeten van een minder vaak gebruikte vreemde taal. Ik hoop te hebben aangetoond dat het lab-onderzoek een veelbelovende weg is voor toekomstig onderzoek naar taalverlies maar ook dat een goede mix van beide methodes belangrijk is als we meer te weten willen komen over hoe we talen vergeten.

Curriculum Vitae

Anne Mickan was born on November 30, 1990, in Dresden, Germany. She obtained her bachelor's degree in English Language and Literature with minors in Spanish and Latin American Studies and Political Science from the Free University of Berlin in 2013. During her undergraduate studies, she spent one year at Reed College in Portland, Oregon, USA, where she took courses in psycholinguistics and psychology. Inspired by these courses she decided to move to Nijmegen, Netherlands, to pursue a master in Cognitive Neuroscience (CNS) at Radboud University with specialization in Language and Communication. She graduated from the CNS program in 2015 (cum laude). Anne then moved on to the Max Planck Institute for Psycholinguistics in Nijmegen, where she joined the Neurobiology of Language department as a research assistant for one year. In 2016, she received an International Max Planck Research School (IMPRS) for Language Sciences fellowship to pursue the PhD work reported on in this thesis. She was supervised by prof. dr. James McQueen and dr. Kristin Lemhöfer. In January 2021, Anne started working as a research software engineer at the Radboud University Medical Center.

Publication List

Mickan, A., McQueen, J. M., & Lemhöfer, K. (2020). Between-language competition as a driving force in foreign language attrition. *Cognition, 198*, 104218. https://doi.org/10.1016/j.cognition.2020.104218

Mickan, A., & Lemhöfer, K. (2020). Tracking syntactic conflict between languages over the course of L2 acquisition: A cross-sectional event-related potentials study. *Journal of Cognitive Neuroscience, 32*(5), 822-846. https://doi.org/10.1162/jocn_a_01528

Mickan, A., McQueen, J. M., & Lemhöfer, K. (2019). Bridging the gap between second language acquisition research and memory science: the case of foreign language attrition. *Frontiers in Human Neuroscience, 13*, 397. https://doi.org/10.3389/fnhum.2019.00397

Mickan, A., Schiefke, M., & Stefanowitsch, A. (2014). Key is a llave is a Schlüssel: A failure to replicate an experiment from Boroditsky et al. 2003. *Yearbook of the Cognitive Linguistics Association. 2*(1), 39-50. https://doi.org/10.1515/gcla-2014-0004

Submitted

Mickan, A., Slesareva, K., McQueen, J., & Lemhöfer, K. (2020). New in, Old out: Does Learning a New Language Make You Forget Previously Learned Foreign Languages? *Manuscript submitted for publication.*

Mickan, A., McQueen, J.M., Valentini, B., Piai, V., & Lemhöfer, K. (2020). Electrophysiological evidence for cross-language interference in foreign-language attrition. *Manuscript submitted for publication.*

Mickan, A., McQueen, J.M., Brehm, L., & Lemhöfer, K. (2020). Individual Differences in Foreign Language Attrition: The Role of Language Use After a Study Abroad. *Manuscript submitted for publication.*

Acknowledgments

Getting this book ready for print feels like a huge achievement. Truth be told, it would not have been possible without the support of a whole lot of people. So here go my thank yous to all the people who contributed to this thesis in one way or another, and without whom I wouldn't have made it this far.

First and foremost, I would like to thank my supervisors, **Kristin** and **James**. Thanks for always making time to discuss puzzling data and brainstorm experimental designs, and thanks for fighting your way through each and every of my usually slightly too long manuscripts and the many confused comments I left in the margins of my Word documents! This book would not exist if it hadn’t been for your tremendous support every step of the way.

Kristin. I have lost track of how many meetings we’ve had and how many emails we’ve exchanged, but I do know that I can count myself very lucky to have had such an involved and supportive supervisor. Thank you for all the time and energy you put into my projects, thank you for all the additional brainstorming meetings that, especially in those last months of thesis writing, really helped me organize my thoughts and pull through; and thank you for putting up with my stubbornness earlier on in the process. Perhaps most importantly though, thank you for always caring about me personally and for being so compassionate, understanding and kind when I was going through difficult times. It meant (and still means) a lot to me.

James. Going into meetings with you, I always knew I’d come out with a feasible and practical solution and a positive mindset. Thanks for always making me look on the bright side, no matter how harsh some of our reviewers’ comments were, or how many extra datasets or experiments they asked for. And thanks for always keeping the PhD timeline in mind and for being pragmatic when it was called for.

I would also like to thank my reading committee, **Rob Schoonen**, **Merel Keijzer** and **Gabriele Janzen**, for taking the time to read and evaluate my dissertation. I’m looking forward to discussing its contents with you during my defense.

For some of the chapters, I had the pleasure to collaborate with and learn from other scientists:

Laurel, I think it's no understatement to say that you're the queen of mixed models. If it hadn't been for your swift replies to my emails and for your help with all things stats related, I would probably still be reading up on contrast coding, convergence warnings and predictor plots. Thanks for your patience and words of wisdom!

Vitória, it was an honor to work with you on Chapter 3. I learned so much about EEG data analysis and science in general from you. Your hands-on supervision style and your incredibly fast responses to all my calls for help made my life so much easier.

Throughout my PhD, I also had the luck to supervise a number of incredibly smart and motivated students. I truly enjoyed working with you and I'm grateful for everything I learned from interacting with each and every one of you. I would also probably still be in the lab if it hadn't been for your help with testing and data preprocessing: **Alicia**, **Beatrice**, **Camille**, **Dennis**, **Doriana**, **Giulio**, **Katya**, **Laura**, **Nikita**, **Orhun** and **Panthea**. The same shout-out goes to **Fabian**, **Iris**, **Julia** and **Katharina** for being the most reliable and fun-to-work with research assistants. I wish you all the best for completing your own PhD journeys.

Collecting the data for this thesis would have been much more cumbersome, if not impossible, if it hadn't been for all the technical support I received over the years. **Gerard** and **Pascal**, thank you for trouble shooting and bug fixing the voice key and button box settings for most of my Presentation scripts with me. **Miriam**, thanks for answering all questions related to research administration. **Arvind** and **Wilbert**, thanks a million for making the online testing for Chapter 5 possible and thank you, **Kees**, **Maurice** and **Cedric**, for creating and maintaining a custom MPI database for participant reimbursement for this study. Without your help, Chapter 5 would have only been a vague idea.

Being an IMPRS-funded PhD student with an office at the DCC was not always easy from an administrative point of view, but thanks to the awesome admin team at the Donders and MPI it was smooth sailing for the most part. Thanks, **Jolanda**, for arranging a 'golden elephant' whenever I needed one. Thanks, **Vanessa**, for dealing with all my special reimbursement requests and sorry for that one time I thought I was being charged for printing costs when really the costs were just my weekly lunch escapes to the university hospital (also thanks to **Mónica** and **Arushi** for their great detective work on this one ;-)). Thanks, **Kevin**, for checking in every once in a while, and for making sure I was doing ok. Thanks also for taking care that I (and everybody else in the IMPRS) acquired much more during the PhD than just the pure research skills.

Sound learning group - thank you for listening to my presentations, for thinking along and for providing thought-provoking feedback that made me see my work from different angles. It was great to be your colleague.

To everyone who was at one point or another part of the **L2 group** meetings – thanks also to you for listening to my occasional presentations and thanks for the many Friday lunches together. **Johanna**, special thanks to you for taking the time to explain your statistical models to me – you saved me a lot of time and trouble! Thanks also for your supportive messages over the last months and your advice on choosing a new job.

The PhD years would have been much greyer without the social contacts; the friends, both close and afar, that supported me along the way.

First of all, I want to thank my paranymphs, **Arushi** and **Mónica**.

Arushi, I'm so so glad that I got to do the PhD with you by my side (literally, for most of the time anyway ;-)). I really couldn't have wished for a smarter, more caring and compassionate, and fiercely critical (in a good way!) office mate. I will (and already do) miss you as my office mate, but I'm glad I won't have to miss you as a friend!

Mónica, my memory of you goes back to our shared time as subeditors for the CNS journal. My hispanophile side was super excited to find out you spoke Spanish. We never actually spoke much Spanish over the years, but we sure did share a lot of sangrias. Thanks for always being there for me when I needed a shoulder to cry on, someone to vent at, someone to listen to Reggeaton with, someone to get a sunburn with, and most importantly someone to laugh with.

Naomi. Drei gemeinsame Jahre in der Johannes Vijghstraat, wöchentliche Lunchdates und regelmäßige Spabesuche haben für jede Menge schöne Erinnerungen gesorgt. Ich bin für jede einzelne dieser Erinnerungen dankbar und schätze es enorm, dass du die vielen Jahre über immer für mich da warst.

Annika, ich denke gern an unsere gemeinsamen Mittagspausen oder unsere gelegentlichen Strandtage zurück und freue mich schon darauf, bald meine outside-of-academia Erfahrungen mit dir zu teilen.

Thanks also to all the other IMPRS and DCC colleagues who brightened up my PhD years in various ways. **Francie**, thanks for inspiring me to 'cook outside the box' every now and then. I hope we continue our unconventional dinner parties! **Sara**, thanks for introducing me to Rladies and for encouraging me to explore options outside of science. **Sophie**, danke für so einige entspannte, spontane Lunchdates und dein stets offenes Ohr. **Jana**, danke für den Napoleon-Plüschersatz.

Egle and **Valentin**. It's pretty special to have friends who move to another city to make hanging out easier. :-P Well, maybe that was not your primary reason to move, but I truly am thankful that we get together so regularly. Thanks for the many Saturdays and Sundays at your place, our place, or some random other place in the Netherlands and the delicious food and fun games that were involved in those hang-outs. In short, thanks for so reliably and regularly distracting me from all things thesis related! **Egle**, special thanks to you for also making sure I stayed fit, even in Covid times, and for the occasional post-workout croissant-and-coffee dates.

Marie, auch wenn wir uns nur selten persönlich gesehen haben, haben wir die Dissertationszeit irgendwie doch gemeinsam durch- und überstanden. Unsere Skype-Gespräche, deine aufmunternden Nachrichten und dein Paket an mich mit dem Wunderkakao für den Endspurt kamen irgendwie immer genau dann, wenn ich sie am dringendsten brauchte.

Thanks also to my other friends from Dresden, Berlin, and around the world for staying in touch all those years and for making the time to come visit me in the Netherlands every now and then: **Caro**, **Claudi**, **Cristina**, **Isabella**, **Kristin**, **Laia**, **Maaike**, **Silke**, **Steffi**, **Olaia**, **Vicky**...

Finally, this book would not have been possible without the tremendous support from my family, both "old" and "new".

Mama und **Papa**, danke, dass ihr immer an mich geglaubt habt und mich immer unterstützt habt, egal wohin meine Reise ging. Trotz der vielen Kilometer, die zwischen uns liegen, seid ihr stets mein erster Anlaufpunkt. Ihr seid immer für mich da und dafür bin euch unendlich dankbar. Mein Dank geht auch an **Karli** und meine lieben Großeltern, **Annerose**, **Hans**, **Gudrun** und **Peter** – danke für den Halt, den ihr mir in meinem Leben gebt. **Opa Hans**, es schmerzt, dass du diese Worte nicht mehr lesen kannst. Ich hoffe, dass du dennoch weißt, was für einen entscheidenden Einfluss du auf mein Leben hattest, nicht zuletzt durch deine ansteckende Leidenschaft für Fremdsprachen.

To my American family, **Nancy** and **Richard**: thank you for everything you did for me while I was in Portland, and thank you even more for continuing to care so deeply, despite all that water between us and the sad fact that we only get to see each other sporadically. Your continued support and your firm belief in me mean a lot to me.

Angela en **Paul**, bedankt dat jullie me in jullie leuke, lieve en gezellige familie hebben opgenomen. Het voelt altijd als vakantie als jullie op bezoek zijn. Ook bedankt aan **Laura** voor de lekkere taartrecepten en de schattige foto's van Clarence, die me vooral in de laatste maanden vaak positief hebben afgeleid. **Paula** en **Harrie**, bedankt voor al het heerlijke eten en de potjes kaarten, die ik nog geen één keer heb gewonnen, maar die me wel elke keer uit mijn werkbubbel hebben gehaald. **Marjo** en **Agaath**, heel erg bedankt voor het lezen en leesbaar maken van de Nederlandse samenvatting.

En ten slotte, mijn lieve **Guido**, ik zou een tweede boek kunnen schrijven over alles waar ik jou dankbaar voor ben, maar laten we het hier kort houden: dank je voor al jouw steun en liefde en dank je dat je me elke dag eraan herinnert dat het leven zoveel meer te bieden heeft dan dit boekje.

MPI Series in Psycholinguistics

1 The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. *Miranda van Turennout*

2 The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. *Niels O. Schiller*

3 Lexical access in the production of ellipsis and pronouns. *Bernadette M. Schmitt*

4 The open-/closed-class distinction in spoken-word recognition. *Alette Haveman*

5 The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. *Kay Behnke*

6 Gesture and speech production. *Jan-Peter de Ruiter*

7 Comparative intonational phonology: English and German. *Esther Grabe*

8 Finiteness in adult and child German. *Ingeborg Lasser*

9 Language input for word discovery. *Joost van de Weijer*

10 Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe. *James Essegbey*

11 Producing past and plural inflections. *Dirk Janssen*

12 Valence and transitivity in Saliba: An Oceanic language of Papua New Guinea. Anna Margetts

13 From speech to words. *Arie van der Lugt*

14 Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language. *Eva Schultze-Berndt*

15 Interpreting indefinites: An experimental study of children's language comprehension. *Irene Krämer*

16 Language-specific listening: The case of phonetic sequences. Andrea Weber

17 Moving eyes and naming objects. *Femke van der Meulen*

18 Analogy in morphology: The selection of linking elements in Dutch compounds. *Andrea Krott*

19 Morphology in speech comprehension. *Kerstin Mauth*

20 Morphological families in the mental lexicon. *Nivja H. de Jong*

21 Fixed expressions and the production of idioms. *Simone A. Sprenger*

22 The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria). *Birgit Hellwig*

23 Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. *Fermín Moscoso del Prado Martín*

24 Contextual influences on spoken-word processing: An electrophysiological approach. *Daniëlle van den Brink*

25 Perceptual relevance of prevoicing in Dutch. *Petra M. van Alphen*

26 Syllables in speech production: Effects of syllable preparation and syllable frequency. *Joana Cholin*

27 Producing complex spoken numerals for time and space. *Marjolein Meeuwissen*

28 Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction. *Rachèl J. J. K. Kemps*

29 At the same time...: The expression of simultaneity in learner varieties. *Barbara Schmiedtová*

30 A grammar of Jalonke argument structure. *Friederike Lüpke*

31 Agrammatic comprehension: An electrophysiological approach. *Marlies Wassenaar*

32 The structure and use of shape-based noun classes in Miraña (North West Amazon). *Frank Seifart*

33 Prosodically-conditioned detail in the recognition of spoken words. *Anne Pier Salverda*

34 Phonetic and lexical processing in a second language. *Mirjam Broersma*

35 Retrieving semantic and syntactic word properties. *Oliver Müller*

36 Lexically-guided perceptual learning in speech processing. *Frank Eisner*

37 Sensitivity to detailed acoustic information in word recognition. *Keren B. Shatzman*

38 The relationship between spoken word production and comprehension. *Rebecca Özdemir*

39 Disfluency: Interrupting speech and gesture. *Mandana Seyfeddinipur*

40 The acquisition of phonological structure: Distinguishing contrastive from non-contrastive variation. *Christiane Dietrich*

41 Cognitive cladistics and the relativity of spatial cognition. *Daniel B.M. Haun*

42 The acquisition of auditory categories. *Martijn Goudbeek*

43 Affix reduction in spoken Dutch. *Mark Pluymaekers*

44 Continuous-speech segmentation at the beginning of language acquisition: Electrophysiological evidence. *Valesca Kooijman*

45 Space and iconicity in German Sign Language (DGS). *Pamela Perniss*

46 On the production of morphologically complex words with special attention to effects of frequency. *Heidrun Bien*

47 Crosslinguistic influence in first and second languages: Convergence in speech and gesture. *Amanda Brown*

48 The acquisition of verb compounding in Mandarin Chinese. *Jidong Chen*

49 Phoneme inventories and patterns of speech sound perception. *Anita Wagner*

50 Lexical processing of morphologically complex words: An information-theoretical perspective. *Victor Kuperman*

51 A grammar of Savosavo, a Papuan language of the Solomon Islands. *Claudia Wegener*

52 Prosodic structure in speech production and perception. *Claudia Kuzla*

53 The acquisition of finiteness by Turkish learners of German and Turkish learners of French: Investigating knowledge of forms and functions in production and comprehension. *Sarah Schimke*

54 Studies on intonation and information structure in child and adult German. *Laura de Ruiter*
55 Processing the fine temporal structure of spoken words. *Eva Reinisch*
56 Semantics and (ir)regular inflection in morphological processing. *Wieke Tabak*
57 Processing strongly reduced forms in casual speech. *Susanne Brouwer*
58 Ambiguous pronoun resolution in L1 and L2 German and Dutch. *Miriam Ellert*
59 Lexical interactions in non-native speech comprehension: Evidence from electro-encephalography, eye-tracking, and functional magnetic resonance imaging. *Ian FitzPatrick*
60 Processing casual speech in native and non-native language. *Annelie Tuinman*
61 Split intransitivity in Rotokas, a Papuan language of Bougainville. *Stuart Robinson*
62 Evidentiality and intersubjectivity in Yurakaré: An interactional account. *Sonja Gipper*
63 The influence of information structure on language comprehension: A neurocognitive perspective. *Lin Wang*
64 The meaning and use of ideophones in Siwu. *Mark Dingemanse*
65 The role of acoustic detail and context in the comprehension of reduced pronunciation variants. *Marco van de Ven*
66 Speech reduction in spontaneous French and Spanish. *Francisco Torreira*
67 The relevance of early word recognition: Insights from the infant brain. *Caroline Junge*
68 Adjusting to different speakers: Extrinsic normalization in vowel perception. *Matthias J. Sjerps*
69 Structuring language. Contributions to the neurocognition of syntax. *Katrien R. Segaert*
70 Infants' appreciation of others' mental states in prelinguistic communication: A second person approach to mindreading. *Birgit Knudsen*
71 Gaze behavior in face-to-face interaction. *Federico Rossano*
72 Sign-spatiality in Kata Kolok: how a village sign language of Bali inscribes its signing space. *Conny de Vos*
73 Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning. *Attila Andics*
74 Lexical processing of foreign-accented speech: Rapid and flexible adaptation. *Marijt Witteman*
75 The use of deictic versus representational gestures in infancy. *Daniel Puccini*
76 Territories of knowledge in Japanese conversation. *Kaoru Hayano*
77 Family and neighbourhood relations in the mental lexicon: A cross-language perspective. *Kimberley Mulder*
78 Contributions of executive control to individual differences in word production. *Zeshu Shao*
79 Hearing speech and seeing speech: Perceptual adjustments in auditory-visual processing. *Patrick van der Zande*

80 High pitches and thick voices: The role of language in space-pitch associations. *Sarah Dolscheid*

81 Seeing what's next: Processing and anticipating language referring to objects. *Joost Rommers*

82 Mental representation and processing of reduced words in casual speech. *Iris Hanique*

83 The many ways listeners adapt to reductions in casual speech. *Katja Poellmann*

84 Contrasting opposite polarity in Germanic and Romance languages: Verum Focus and affirmative particles in native speakers and advanced L2 learners. *Giuseppina Turco*

85 Morphological processing in younger and older people: Evidence for flexible dual-route access. *Jana Reifegerste*

86 Semantic and syntactic constraints on the production of subject-verb agreement. *Alma Veenstra*

87 The acquisition of morphophonological alternations across languages. *Helen Buckler*

88 The evolutionary dynamics of motion event encoding. *Annemarie Verkerk*

89 Rediscovering a forgotten language. *Jiyoun Choi*

90 The road to native listening: Language-general perception, language-specific input. *Sho Tsuji*

91 Infants'understanding of communication as participants and observers. *Gudmundur Bjarki Thorgrímsson*

92 Information structure in Avatime. *Saskia van Putten*

93 Switch reference in Whitesands. *Jeremy Hammond*

94 Machine learning for gesture recognition from videos. *Binyam Gebrekidan Gebre*

95 Acquisition of spatial language by signing and speaking children: a comparison of Turkish sign language (TID) and Turkish. *Beyza Sümer*

96 An ear for pitch: on the effects of experience and aptitude in processing pitch in language and music. *Salomi Savvatia Asaridou*

97 Incrementality and Flexibility in Sentence Production. *Maartje van de Velde*

98 Social learning dynamics in chimpanzees: Reflections on (nonhuman) animal culture. *Edwin van Leeuwen*

99 The request system in Italian interaction. *Giovanni Rossi*

100 Timing turns in conversation: A temporal preparation account. *Lilla Magyari*

101 Assessing birth language memory in young adoptees. *Wencui Zhou*

102 A social and neurobiological approach to pointing in speech and gesture. *David Peeters*

103 Investigating the genetic basis of reading and language skills. *Alessandro Gialluisi*

104 Conversation Electrified: The Electrophysiology of Spoken Speech Act Recognition. *Rósa Signý Gisladottir*

105 Modelling Multimodal Language Processing. *Alastair Smith*

106 Predicting language in different contexts: The nature and limits of mechanisms in anticipatory language processing. *Florian Hintz*

107 Situational variation in non-native communication. *Huib Kouwenhoven*

108 Sustained attention in language production. *Suzanne Jongman*

109 Acoustic reduction in spoken-word processing: Distributional, syntactic, morphosyntatic, and orthographic effects. *Malte Viebahn*

110 Nativeness, dominance, and the flexibility of listening to spoken language. *Laurence Bruggeman*

111 Semantic specificity of perception verbs in Maniq. *Ewelina Wnuk*

112 On the identification of FOXP2 gene enhancers and their role in brain development. *Martin Becker*

113 Events in language and thought: The case of serial verb constructions in Avatime. *Rebecca Defina*

114 Deciphering common and rare genetic effects on reading ability. *Amaia Carrión Castillo*

115 Music and language comprehension in the brain. *Richard Kunert*

116 Comprehending Comprehension: Insights from neuronal oscillations on the neuronal basis of language. *Nietzsche H.L. Lam*

117 The biology of variation in anatomical brain asymmetries. *Tulio Guadalupe*

118 Language processing in a conversation context. *Lotte Schoot*

119 Achieving mutual understanding in Argentine Sign Language. *Elizabeth Manrique*

120 Talking Sense: the behavioural and neural correlates of sound symbolism. *Gwilym Lockwood*

121 Getting under your skin: The role of perspective and simulation of experience in narrative comprehension. *Franziska Hartung*

122 Sensorimotor experience in speech perception. *Will Schuerman*

123 Explorations of beta-band neural oscillations during language comprehension: Sentence processing and beyond. *Ashley Lewis*

124 Influences on the magnitude of syntactic priming. *Evelien Heyselaar*

125 Lapse organization in interaction. *Elliott Hoey*

126 The processing of reduced word pronunciation variants by natives and foreign language learners: Evidence from French casual speech. *Sophie Brand*

127 The neighbors will tell you what to expect: Effects of aging and predictability on language processing. *Cornelia Moers*

128 The role of voice and word order in incremental sentence processing. *Sebastian Sauppe*

129 Learning from the (un)expected: Age and individual differences in statistical learning and perceptual learning in speech. *Thordis Neger*

130 Mental representations of Dutch regular morphologically complex neologisms. *Laura de Vaan*

131 Speech production, perception, and input of simultaneous bilingual preschoolers: Evidence from voice onset time. *Antje Stoehr*

132 A holistic approach to understanding pre-history. *Vishnupriya Kolipakam*

133 Characterization of transcription factors in monogenic disorders of speech and language. *Sara Busquets Estruch*

134 Indirect request comprehension in different contexts. *Johanne Tromp*

135 Envisioning Language - An Exploration of Perceptual Processes in Language Comprehension. *Markus Ostarek*

136 Listening for the WHAT and the HOW: Older adults' processing of semantic and affective information in speech. *Juliane Kirsch*

137 Let the agents do the talking: on the influence of vocal tract anatomy on speech during ontogeny and glossogeny. *Rick Janssen*

138 Age and hearing loss effects on speech processing. *Xaver Koch*

139 Vocabulary knowledge and learning: Individual differences in adult native speakers. *Nina Mainz*

140 The face in face-to-face communication: Signals of understanding and non-understanding. *Paul Hömke*

141 Person reference and interaction in Umpila/Kuuku Ya'u narrative. *Clair Hill*

142 Beyond the language given: The neurobiological infrastructure for pragmatic inferencing. *Jana Bašnáková*

143 From Kawapanan to Shawi: Topics in language variation and change. *Luis Miguel Rojas-Berscia*

144 On the oscillatory dynamics underlying speech-gesture integration in clear and adverse listening conditions. *Linda Drijvers*

145 Understanding temporal overlap between production and comprehension. *Amie Fairs*

146 The role of exemplars in speech comprehension. *Annika Nijveld*

147 A network of interacting proteins disrupted in language-related disorders. *Elliot Sollis*

148 Fast speech can sound slow: Effects of contextual speech rate on word recognition. *Merel Maslowski*

149 Reason-giving in everyday activities. *Julija Baranova*

150 Speech planning in dialogue - Psycholinguistic studies of the timing of turn taking. *Mathias Barthel*

151 The role of neural feedback in language unification: How awareness affects combinatorial processing. *Valeria Mongelli*

152 Exploring social biases in language processing. *Sara Iacozza*

153 Vocal learning in the pale spear-nosed bat, Phyllostomus discolor. *Ella Lattenkamp*

154 The effect of language contact on speech and gesture: The case of Turkish-Dutch bilinguals in the Netherlands. *Elif Zeynep Azar*

155 Language and society: How social pressures shape grammatical structure. *Limor Raviv*

156 The moment in between: Planning speech while listening. *Svetlana-Lito Gerakaki*

157 How speaking fast is like running: Modelling control of speaking rate. *Joe Rodd*

158 The power of context: How linguistic contextual information shapes brain dynamics during sentence processing. *René Terporten*

159 Neurobiological models of sentence processing. *Marvin Uhlmann*

160 Individual differences in syntactic knowledge and processing: The role of literacy experience. *Saoradh Favier*

161 Memory for speaking and listening. *Eirini Zormpa*

162 Masculine generic pronouns: Investigating the processing of an unintended gender cue. *Theresa Redl*

163 Properties, structures and operations: Studies on language processing in the brain using computational linguistics and naturalistic stimuli. *Alessandro Lopopolo*

164 Investigating spoken language comprehension as perceptual inference. *Greta Kaufeld*

165 What was that Spanish word again? Investigations into the cognitive mechanisms underlying foreign language attrition. *Anne Mickan*

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

VISITING ADDRESS
Wundtlaan 1
6525 XD Nijmegen
The Netherlands

POSTAL ADDRESS
P.O. Box 310
6500 AH Nijmegen
The Netherlands

CONTACT
T +31(0)24 3521 911
F +31(0)24 3521 213
E info@mpi.nl
Twitter @MPI_NL
www.mpi.nl

DONDERS INSTITUTE