### 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.

### 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 (Choi, Broersma, & Cutler, 2017; Isurin, 2000; Pallier et al., 2003; Pierce, Klein, Chen, Delcenserie, & Genesee, 2015; 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 & Phelphs, 1987; Bahrick, 1984; de Bot & Weltens, 1995; Mehotcheva, 2010; Murtagh, 2003; Weltens, Van Els, & Schils, 1989; Grendel, 1993; Tomiyama, 2008; Weltens, 1988; Xu, 2010; for an overview, see Schmid & Mehotcheva, 2012; Mehotcheva & Köpke, 2019). A seminal case study by Bahrick (1984) 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, 1984; de Groot & 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., Olshtain, 1989; Kuhberg, 1992). 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 et al., 1989), while others find notable forgetting already after just a year or even less (e.g., Bahrick, 1984; 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 factors that are believed to impact individual forgetting rates. Among those are proficiency at attrition onset (e.g., Bahrick, 1984; Mehotcheva, 2010; Woltens, 1988), age at attrition onset (e.g., Bardovi-Harlig & Stringer, 2010), as well as language usage patterns and motivation to maintain the foreign language (e.g., Mehotcheva, 2010). For a recent discussion of extra-linguistic factors in FL attrition, see Mehotcheva and 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, Martens, & Stoessel, 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 after about a day. 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 (McGeech, 1932). 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 this latter notion, interference theory emerged. Rather than attributing forgetting to the mere passage of time and disuse of information, interference theory assumes forgetting to be a 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; McGeech, 1932). 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, Bjork, & Bjork, 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, Aslan, Bäuml, Gäbel, & Mecklinger, 2007; Penolazzi, Stramaccia, Braga, Mondini, & Galfano, 2014). The RIF effect has been replicated using a great variety of stimuli materials, including not only category-exemplar pairs in verbal lists, but also pictures (Ford, Keating, & Patel, 2004; Hulbert & Norman, 2014; Reppa, Williams, Worth, Greville, & Saunders, 2017), faces (Wimber, Alink, Charest, Kriegeskorte, & Anderson, 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 more specifically the forgetting of FL vocabulary. In line with Linck and Kroll (2019) and Mickan, McQueen, and Lemböfer (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. 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, Caramazza, & Sebastián-Gallés, 2000). Mostly though, 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, Montoya, & Werner, 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, Santesteban, & Ivanova, 2006; Linck, Hoshino, & Kroll, 2008). Although it has been argued that the presence of competition effects and thus the need for inhibition largely depends 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 compete for selection (for recent reviews see Kroll, Bobb, Misra, & Guo, 2008, and Kroll, Gullifer, & Rossi, 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, McVeigh, Marful, and Anderson (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 min, 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 also 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 aim to address the above studies’ shortcomings. We assess 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) are asked to retrieve half of the recently learned words in another language, and a final L3 Spanish test phase. We hypothesize that the retrieval of translation equivalents will 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 are separated by a night’s sleep to allow for consolidation of the newly learned L3 words, and interference and final test are 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 asks 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, Forster, & Frost, 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, Schellenberger, and Schriefers (2018) 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. Experiment 1

2.1. Methods

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 Item selection below), 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 (mean age = 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 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 they had 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 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 Procedure below) (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.
Interference condition, and subsequently averaged over groups. Ranges show the absolute min and max values per group and condition.

Note. Itemsets differed across participants, as described in the Item selection section. Means (N=23)

**Table 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 |
| LoI (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  |
| **Dutch interference group** |      |       |       |      |       |       |
| AoA                   | 9.96  | 2.03  | 5–13  | 10.20 | 1.24  | 7–12  |
| LoI (years)           | 11.39 | 5.21  | 6–26  | 11.50 | 3.40  | 7–20  |

Note. $M =$ mean; $SD =$ standard deviation; AoA = age of acquisition; LoI = length of exposure; FAR = false alarm rate (in %).

**2.1.2. Materials**

The stimulus database consisted of 169 Spanish nouns referring to concrete, everyday objects or animals (see Appendix A 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, Piepenbrock, & Gulikers, 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, Dionne-Dostie, Montreuil, & Lepage, 2010). Photos were all embedded in a 6 × 6 cm white frame with the depicted object/animal centered and adjusted for size to occupy a maximum of 400px in either width or length. Furthermore, each noun was recorded by a Spanish native speaker from Madrid (Spain).

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 pretest 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 pretest, see ‘Pretest’ section for details) were already known to a participant, a Matlab script subsequently replaced those items with unknown words from the remaining pretest items (mean words replaced = 6.16, range = 0–24, see Supplementary materials, S1, 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 (see Supplementary material for details). 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, Keuleers, & Brysbaert, 2017), as well as phonological similarity in Spanish assessed via Levenshtein distances (Levenshtein, 1966) (see Table 2 for averages). Word frequency was not explicitly controlled for given the amount of constraints we already had. However, as Table 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 Supplementary materials, S2).

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, www.neurobs.com). The computer screen (BenQ XL 2420Z, 23-in.) was set to white, with a resolution of 1920 × 1080 pixels at a refresh rate of 60 Hz. All audio stimuli were

**Table 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  | 0.78  | 2–5   |
| Dutch Celex 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 Celex per million frequency | 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 | 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 |
| **Spanish Levenshtein distance** | 6.38  | 1.45  | 0–7   | 6.14  | 1.45  | 0–7   | 6.41  | 1.45  | 0–7   | 6.41  | 1.45  | 0–7   |
| Across-set semantic distance | 0.82  | 0.09  | 0.31–1.09 | 0.82 | 0.09  | 0.31–1.09 | 0.82  | 0.09  | 0.31–1.09 | 0.82  | 0.09  | 0.31–1.09 |

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.

* For an explanation on how we assessed semantic similarity, please refer to the Supplementary materials (Supplementary materials, S1).
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 Fig. 1): On day 1, participants learned 40 new Spanish words in a mixture of recognition and production tasks. On day 2, 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.

2.1.3.1. Day 1 – L3 Spanish learning phase

2.1.3.1.1. Pretest. 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 2 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 Supplementary materials, S1, for details). To make this pretest 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.

2.1.3.1.2. Learning tasks. The learning phase consisted of a mix of comprehension and production tasks (see Fig. 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 rows in a round. Feedback was provided on all learning tasks (see details for each task separately). For inter-stimulus intervals and feedback timing please consult the Supplementary materials (Supplementary materials, S3).

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 Supplementary materials, S4, 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 30s 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 2 h. The adaptive learning task was stopped after 1 h 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).

### 2.1.3.2. Day 2 – L1/L2 interference phase & L3 Spanish posttest.

One day after the learning session (Dutch group: $M = 24.03$ h, $SD = 2.89$, range = $18–29$, English group: $M = 24.13$ h, $SD = 2.41$, 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 Fig. 3). In the familiarization round, participants saw the picture together with the first syllable of the English or Dutch word (the first grapheme for

| Interference Phase |
|--------------------|
| **Word completion in L1 or L2** |
| 1 round, no feedback |
| **Picture Naming in L1 or L2** |
| (for half the items) |
| L1 or L2 |
| 2 rounds, no feedback |
| **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** |
| 2 rounds, no feedback |
| **Letter search in L1 or L2** |
| 2 rounds, no feedback |

*Fig. 3. Overview of interference phase tasks.*
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 pretest 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 pretest 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.

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.

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 Supplementary materials, S3, for task design details). Together they took roughly 20 min. 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 1. Since these tasks merely served as filler tasks, we did not analyze them any further.

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.

2.1.3.3. 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 (Lemböfer & Broersma, 2012; see Table 1 for group means).

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 Supplementary materials, S4, 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, Schriefers, and Lemböfer (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 3 exemplifies the scoring procedure for the ‘embudo’ example.

‘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 Modelling 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: $(1 / (5 + 1)) \times 100 = 16.67\%$.

2.1.5. Reaction time measurements

Naming latencies were measured 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 participant 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.

2.1.6. Modelling

We analyzed the data using generalized mixed effects models with the lme4 package (version 1.1-15, Bates, Mächler, Bolker, & Walker, 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 probability (i.e., for the ‘embuda’ example above, the probability of correctly producing 5 out of 6 phonemes), 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 (2 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, Levy, Scheepers, & Tily, 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, Jackson, & Miller, 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 posttest. For analyses on the raw naming latencies at final test, rather than difference scores, see the supplementary materials (Supplementary materials, S5). These analyses lead to the same conclusions as the primary analyses based on difference scores.
2.2. Results

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).

2.2.2. Naming performance after interference (on days 2 and 8)

2.2.2.1. Naming accuracy
Fig. 4 shows the mean percentages of incorrectly recalled phonemes per target word 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 4 and 5.

![Fig. 4. Experiment 1. Error rates in Spanish productions as measured in percentage of incorrectly recalled phonemes per target word for the final tests on day 2 and 8 respectively. Error bars reflect standard error around the condition means.](image)

**Table 3**
Scoring example, phonetically transcribed.

| Target word | ɛ̃ | m | b | u | δ | ɑ |
|-------------|----|---|---|---|---|---|
| Participants production | ɛ̃ | m | b | u | δ | ɑ |
| Scoring | Correct | Correct | Correct | Correct | Correct | Incorrect (substitution) |

**Table 4**
Mixed effect model output for naming accuracy in Experiment 1.

| Fixed effects | Estimate | SE  | z     | p(χ²)  |
|---------------|----------|-----|-------|--------|
| 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 × Language × Day |  0.73 |  0.27|  2.67 | .011   |

| 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|
|                | Interference × Day | 1.53| 1.24| 0.21| −0.21| 0.06|

Note. Significant effects are marked in bold. SE = standard error; SD = standard deviation; p(χ²) = Chi-square test statistic; Var = variance; Corr = correlation.
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 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.

### 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 Fig. 5 and model outcomes are shown in Tables 6 and 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

---

**Table 5**

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

| Fixed effects                  | Estimate | SE  | z   | p(χ²) | Estimate | SE  | z   | p(χ²) |
|--------------------------------|----------|-----|-----|-------|----------|-----|-----|-------|
| Intercept                      | 4.92     | 0.38| 12.96| <.001 | 3.80     | 0.34| 11.10| <.001 |
| Interference                   | −0.98    | 0.29| −4.33| .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| 1.678|     |      |
| Interference     | 2.26   | 1.50| 0.13|      | Interference 1.72 | 1.31 | 0.33|      |

Note. Significant effects are marked in bold. SE = standard error; SD = standard deviation; p(χ²) = Chi-square test statistic; Var = variance; Corr = correlation.

**Table 6**

Mixed effects model output for naming latencies in Experiment 1.

| Fixed effects                  | Estimate | SE  | t    | p(χ²) |
|--------------------------------|----------|-----|------|-------|
| 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| .792  |
| 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; SD = standard deviation; p(χ²) = Chi-square test statistic; Var = variance; Corr = correlation.

---

Fig. 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 standard error around the condition means.
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; Báuml, Zellner, & Vilimek, 2005; 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 online 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). More importantly, 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 thus 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, Cardoso, & Collins, 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 (De Angelis, 2005; 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 neuroscience 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

### Table 7

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

| Fixed effects | Model output for Day 2 |  |  |  | Model output for Day 8 |  |  |  |
|---------------|------------------------|---|---|---|------------------------|---|---|---|
|               | Estimate | SE  | t  | p(χ²) | Estimate | SE  | t  | p(χ²) |
| 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  |

Note. Significant effects are marked in bold. SE = standard error; SD = standard deviation; p(χ²) = Chi-square test statistic; Var = variance.

slower on day 8 (738 ms slower than at baseline) as compared to day 2 (336 ms slower than at baseline).

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 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).

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 was 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; Báuml, Zellner, & Vilimek, 2005; 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 online 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). More importantly, 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 thus 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, Cardoso, & Collins, 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 (De Angelis, 2005; 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

A. Mickan, et al.
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, Cowell, and Varley (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, 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 L2 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 attempt first of all to replicate the main effect of interference reported in Experiment 1. We also test 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 manipulate 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 manipulate 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 receive interference in their mother tongue Dutch. However, for one group, the interferers are high frequency Dutch words (resembling the L1 interference condition in Experiment 1), while the other group receives 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 expect to replicate the main effect of interference observed in Experiment 1.

3. Experiment 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.

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 (mean age = 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 8. In contrast to Experiment 1, participants

| Table 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    | 2.14    | 6–17   | 11.00   | 3.00    | 6–20   |
| English frequency of use (min/day) |          |         |        |         |         |        |
| - Speaking    | 79      | 247     | 0–1200 | 18      | 25      | 0–60   |
| - Listening   | 165     | 232     | 4–1200 | 178     | 162     | 60–720 |
| - Reading     | 143     | 194     | 0–400  | 91      | 91      | 0–300  |
| - Writing     | 53      | 102     | 0–480  | 33      | 57      | 0–240  |
| English proficiencya |          |         |        |         |         |        |
| - 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 taskb   | 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; FAR = false alarm rate (in %).

a Proficiency self-ratings were given on a scale from 1 (very bad)–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.

"Note that the maximum of 1200 min is an outlier in the dataset – surely this participant does not speak English for an average of 20 h a day. It is possible that he/she either by accident typed in the wrong number, or that they misunderstood the question."
Table 9
Characteristics of items used in Experiment 2.

|                      | High frequency                      | Low frequency                      |
|----------------------|-------------------------------------|------------------------------------|
|                      | Set 1                               | Set 2                               |
|                      | M  SD  Range                        | M  SD  Range                        |
| Spanish word length  | 2.75 0.78 2-5                       | 2.60 0.75 2-4                       |
| (in syllables)       |                                    |                                    |
| Dutch Celex log       | 1.61 0.26 1.23-2.26                | 1.75 0.50 0.95-2.95                |
| frequency            |                                    |                                    |
| Dutch Celex per       | 50.70 45.13 17-180                 | 115.3 197.13 9-900                 |
| million frequency     |                                    |                                    |
| Within-set semantic   | 0.75 0.20 0-1.01                    | 0.74 0.19 0-1.02                    |
| distance a            |                                    |                                    |
| Spanish Levenshtein   | 5.71 1.35 2-9                       | 6.25 1.40 2-10                      |
| distance              |                                    |                                    |
| Across-set semantic   | 0.82 0.12 0.49-1.17                 | 0.79 0.10 0.44-1.03                 |
| distance b            |                                    |                                    |

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.

3.1.3. Procedure

As in Experiment 1, a day after the learning session, participants returned for the interference session (High frequency group: $M = 23.86h$, $SD = 2.79$, 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.

3.1.4. Modelling

Responses and naming latencies were scored exactly as in Experiment 1, and were also analyzed using the same (generalized) mixed-effect 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: no interference, interference), Frequency (2 levels: high frequency, low frequency) and their interaction. Both were again effects coded (-0.5, 0.5). 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. 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.

3.2. Results

3.2.1. Learning success on Day 1

As in Experiment 1, participants were very successful at learning the 3 Please see the Supplementary materials for analyses based on the raw naming latencies (Supplementary materials, S5). The analyses on raw latencies lead to the same conclusions as those based on differences scores.
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).

3.2.2. Naming performance after interference (on day 2)

3.2.2.1. Naming accuracy

Fig. 6 shows the mean percentages of correctly recalled phonemes per target word 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 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.

3.2.2.2. Naming latencies

Naming latencies (in ms) at final test on day 2 per Interference and Frequency condition are plotted in Fig. 6 and model outcomes are shown in Table 10. As in Experiment 1, naming latencies refer to the difference in naming speed between baseline (immediately after learning) and 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 (732ms) than for words that had not (210ms). There was no main effect of frequency.

| Fixed effects | Naming accuracy | Naming latencies |
|---------------|-----------------|-----------------|
| Intercept     | 4.31            | 0.24            |
| SE            | 0.27            | 0.03            |
| z             | 15.82           | 7.01            |
| p(χ²)         | <.001           | <.001           |
| Interference  | −1.38           | 0.25            |
| SE            | 0.25            | 0.03            |
| z             | −5.47           | 7.67            |
| p(χ²)         | <.001           | <.001           |
| Frequency     | −0.07           | 0.02            |
| SE            | 0.53            | 0.07            |
| z             | −0.13           | 0.34            |
| p(χ²)         | .894            | .728            |
| Interference × Frequency | −0.91 | 0.11 |
| SE            | 0.44            | 0.07            |
| z             | −2.04           | 1.60            |
| p(χ²)         | .057            | .106            |

Table 10 Mixed effects model outcome for naming accuracy and latencies in Experiment 2.

| Random effects | Naming accuracy | Naming latencies |
|----------------|-----------------|-----------------|
| Groups         | Var SD Corr     | Groups Var SD Corr |
| Intercept      | 1.66 1.29 −0.38 | Intercept 0.00 0.05 |
| Subject        | 1.88 1.37       | Intercept 0.04 0.21 |
| Interference   | 1.39 1.18       | Interference 0.02 0.51 0.26 |

Note. Significant effects are marked in bold. SE = standard error; SD = standard deviation; p(χ²) = Chi-square test statistic; Var = variance; Corr = correlation.
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.

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.

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 in L3 Spanish 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.

These main effects 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; Bäuml et al., 2005; Johansson et al., 2007; 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 resulting between-language competition, are well-known and thoroughly studied phenomena in psycholinguistics (Kroll et al., 2008). 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 words in other languages, and crucially that these effects persist well beyond the single trial (and thus differ from typical language switch costs; Costa & Santesteban, 2004; Declerck & Philipp, 2015; Meuter & Allport, 1999; and blocked switch costs; Declerck & Philipp, 2017) and that they survive (at the least) a 20-minute delay (not tested in Levy et al., 2007, or 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 min, 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.

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.4

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.

4 Our work is further in line with cue-based models of forgetting, and more generally speaking lends support to models of short and long-term memory that stress the relative diagnostic value of retrieval cues for memory performance (for further discussion, see Naime, 2002).

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 (Bailey & Newman, 2018; Gómez-Ariza, Lechuga, & Pelegri, 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 languages switching tasks (e.g., Costa, La Heij, & Navarrete, 2006; Costa, Miozzo, & Caramazza, 1999; Kroll et al., 2008; Kroll, Dussias, Bogulska, & Kroff, 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 presented in the Introduction that forgetting is not an ‘all or nothing’ phenomenon, but instead a gradual process described by changes in accessibility over time. 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.

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 min, 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) processes in long-term memory. Next to showing that interference effects persist for at least 20 min, 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) processes in long-term memory.

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, Kroll, & Bajo, 2013), but it can also have negative consequences in situations where previously irrelevant, inhibited material suddenly becomes relevant again (see Treccani, Argyri, Sorace, & Della Sala, 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 more 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 Supplementary materials, S6, 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). It would thus be very interesting 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 interference. 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 Supplementary materials, S6, 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 Supplementary materials, S6). 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 differ-
ences, as well as of course the conclusions we draw on the basis of
Experiment 2 in the present paper.

5. Conclusions

The experiments reported in this paper 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 ef-
fects are strongest between foreign languages. Finally, we show that
between-language interference effects are not just momentary forget-
ting 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.

Author contribution statement

Anne Mickan: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing –
Original Draft, Writing – Review & Editing, Visualization, Project administration.
James M. McQueen: Conceptualization, Methodology, Writing –
Review & Editing, Supervision.
Kristin Lemhöfer: Conceptualization, Methodology, Writing –
Review & Editing, Supervision.

Acknowledgements

We would 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 and Julia Egger and Iris Verpaalen for help
with data acquisition. We also thank Michael Masson and Marc
Brysbaert for helpful comments on an earlier version of this manuscript.
This work was supported by an International Max Planck Research
School for Language Sciences fellowship (grant period: 2016–2020).
The funding source had no involvement in any stage of the preparation
of this manuscript. The data can be accessed via the following URL:
http://hdl.handle.net/11633/aabpqtcv.

Declaration of competing interest

None.

Appendix A. Stimulus database for Experiment 1 with translations in English, Dutch and Spanish

| English  | Dutch    | Spanish   |
|----------|----------|-----------|
| iron     | strikijzer| plancha   |
| carrot   | wortel   | zanahoria |
| bush     | struik   | arbusto   |
| needle   | naald    | aguja     |
| swing    | schommel | columpio  |
| dice     | dobbelsteen| dado     |
| mushroom | paddenstoel| hongo    |
| root     | wortel   | raiz      |
| 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    | frengona  |
| 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  |
| straw    | rietje   | paja      |
| butterfly| vlinder  | mariposa  |
| egg      | ei       | huevo     |
| mirror   | spiegel  | espejo    |
| key      | sleutel  | llave     |
| dress    | jurk     | vestido   |
| English                  | French                  | Dutch                  |
|-------------------------|-------------------------|------------------------|
| painting                | schilderij              | cuadro                 |
| cow                     | koe                     | vaca                   |
| spoon                   | lepel                   | cucharra               |
| strawberry              | aardbei                 | fresa                  |
| toothbrush              | tandenborstel          | cepillo de dientes     |
| knife                   | mes                     | cuchillo               |
| glasses                 | bril                    | gafas                  |
| watch                   | horloge                 | reloj                  |
| candle                  | kaars                   | vela                   |
| onion                   | ui                      | eebolla                |
| 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            | stolplicht              | semáforo               |
| rope                    | touw                    | soga                   |
| frog                    | kikker                  | rana                   |
| tie                     | stropdas                | corbata                |
| bucket                  | emmer                   | cubo                   |
| shell                   | schelp                  | concha                 |
| backpack                | rugzak                  | mochila                |
| snail                   | slak                    | carasol                |
| 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                      | bil                     | hacha                  |
| mower                   | grasmaaier              | cortacésped            |
| jar                      | pot                     | tarro                  |
| paintbrush              | kwast                   | pincele                |
| basket                  | mand                    | canasta                |
| saw                     | zaag                    | sierra                 |
| coat                    | jas                     | abrigo                 |
| headphones              | koptelefoon              | cascos                 |
| match                   | lucifer                 | orejillo               |
| razor                   | scheermes               | cuchilla               |
| slide                   | glijbaan                | tobogán                |
| screwdriver             | schroevedraaier         | destornillador         |
| screw                   | schroef                 | navaja                 |
| hairdryer               | fühl                     | 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                  | niemachines              | grapadora              |
| pencilsharpener         | puntenslijper           | sacapuntas             |
| English      | Dutch     | Spanish   | Item type          |
|-------------|-----------|-----------|--------------------|
| back        | rug       | espalda   | Experimental       |
| leg         | been     | pierna    | Experimental       |
| mirror      | spiegel   | espejo    | Experimental       |
| coat        | jas       | abrigo    | Experimental       |
| mug         | kopje     | taza      | Experimental       |
| dress       | jurk      | vestido   | Experimental       |
| knife       | mes       | cuchillo  | Experimental       |
| painting    | schilderij| cuadro    | Experimental       |
| ball        | bal       | pelota    | Experimental       |
| pot         | pan       | olla      | Experimental       |
| watch       | horloge   | reloj     | Experimental       |
| duck        | eend      | pato      | Experimental       |
| candle      | kaars     | vela      | Experimental       |
| fence       | hek       | alambrada | Experimental       |
| carrot      | wortel    | zanahoria | Experimental       |
| suit        | pak       | traje     | Experimental       |
| ring        | ring      | anillo    | Experimental       |
| bell        | bel       | campana   | Experimental       |
| gift        | cadeau    | regalo    | Experimental       |
| moustache   | snor      | bigote    | Experimental       |
| table       | tafel     | mesa      | Experimental       |
| window      | raam      | ventana   | Experimental       |
| horse       | paard     | caballo   | Experimental       |
| nose        | neus      | nariz     | Experimental       |
| egg         | ei        | huevo     | Experimental       |

Note. Items are in the order as presented in the pretest.

Appendix B. Stimulus list in Experiment 2 with translations in English, Dutch and Spanish
suitcase
cloud
wardrobe
steering wheel
pig
belt
branch
blanket
bush
cage
ashtray
lock
fly
woman
snake
book
bread
stone
eye
plate
cheese
bicycle
glasses
bucket
lid
feather
garlic
sausage
bottle
bone
wood
bird
tooth
bowl

Low frequency items
whisk
skunk
apron
funnel
blackberry
nut
cage

Cognition 198 (2020) 104218
Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2020.104218.

References

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language, 49*, 415–445. https://doi.org/10.1016/S0749-596X(02)00006-9.

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*, 1063–1087. https://doi.org/10.1037/0278-7393.20.5.1063.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. *The CELEX lexical database.* (1995). Retrieved from http://celex.mpdl.mpg.de (Database).

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

Bahlbirk, H. P., & Phelphs, E. (1987). Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 344–349. https://doi.org/10.1037/0278-7393.13.3.344.

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*.

Bardel, C., & Falck, Y. (2007). The role of the second language in third language acquisition: The case of Germanic syntax. *Second Language Research, 23*, 459–484. https://doi.org/10.1076/0267658077805577.

Bardovi-Harlig, K., & Stringer, D. (2010). Variables in second language attrition. *Studies in Second Language Acquisition, 32*, 1–45. https://doi.org/10.1017/S0277530210000246.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for *lme4: Linear mixed-effects models confirmatory hypothesis testing: Keep it maximal.* R package devuelvendo, version 0.3-2. https://cran.r-project.org/web/packages/lme4/index.html.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models.* https://cran.r-project.org/web/packages/lme4/index.html.

Bauck, L. A., & Wulff, K. (2007). 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*, 1057–1074. https://doi.org/10.1037/0278-7993.32.5.1057.

Bayley, M. J. (2007). *The R book.* Chichester, UK: John Wiley & Sons. https://doi.org/10.1002/9781118489088.

Bedejas, G. (2005). Interlanguage transfer of function words. *Language Learning, 55*, 379–414. https://doi.org/10.1111/j.1056-386x.2005.00310.x.

de Bot, K., Martens, V., & Stoessel, S. (2004). Finding residual lexical knowledge: The “savings” approach to testing vocabulary. *International Journal of Bilingualism, 8*, 273–382. https://doi.org/10.1177/13670069040031101.

de Bot, K., & Weltens, B. (1995). Foreign language attrition. *Annual Review of Applied Linguistics, 15*, 151–164. https://doi.org/10.1016/S02671919(05)00266X.

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*, 1–56. https://doi.org/10.1111/0023-8333.01119.

de Vos, J. F., Schrieffers, H., & Lehmérof, K. (2018). Noticing vocabulary holes aids in incidental second language word learning: An experimental study. *Bilingualism: Language and Cognition, 22*, 500–515. https://doi.org/10.1017/s1366728918000019.

Declerck, M., & Philipp, A. M. (2015). A review of control processes and their locus in language switching. *Psychonomic Bulletin and Review, 22*, 1630–1645. https://doi.org/10.3758/s13423-015-0836-1.

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*, 488–493. https://doi.org/10.1080/22399140.2016.1250928.

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

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

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology.* New York: Teachers College, Columbia University (H. A. Ruger & C. E. Bussenius, Trans.).

Ecke, P. (2004). Language attrition and theories of forgetting: A cross-disciplinary review. *International Journal of Bilingualism, 8*, 321–354. https://doi.org/10.1177/13670069040030901.

Ford, R. M., Keating, S., & Patel, R. (2004). Retrieval-induced forgetting: A developmental study. *British Journal of Developmental Psychology, 22*, 583–605. https://doi.org/10.1348/0261540042200389.

Garcia-Bajos, E., Migueles, M., & Anderson, M. C. (2009). Script knowledge modulates retrieval-induced forgetting for eyewitness events. *Memory, 17*, 92–103. https://doi.org/10.1080/09658210802372454.

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*, 1122–1139. https://doi.org/10.1037/0278-7393.23.5.1122.

Gollan, T. H., Montoya, R. I., & Werner, G. A. (2000). Semantic and letter fluency in Hebrew–English bilinguals. *Quarterly Journal of Experimental Psychology, 53*, 831–852. https://doi.org/10.1080/027249810.2016.1250928.

Gómez-Ariza, C. J., Lechuga, M. T., & Pellegrina, S. (2005). Retrieval-induced forgetting in recall and recognition of thematically related and unrelated sentences. *Memory & Cognition, 33*, 1431–1441. https://doi.org/10.3758/BF03193376.
Williams, C. C., & Zacks, R. T. (2001). Is retrieval-induced forgetting an inhibitory process? American Journal of Psychology, 114, 329–354. https://doi.org/10.2307/1423685.

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, 582–589. https://doi.org/10.1038/nn.3973.

Xu, X. (2010). English language attrition and retention in Chinese and Dutch university students. University of Groningen (Unpublished doctoral dissertation).

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, 205–216. https://doi.org/10.1017/S0272263100000619.

Williams, S., & Hammarberg, B. (1998). Language switches in L3 production: Implications for a polyglot speaking model. Applied Linguistics, 19, 295–333. https://doi.org/10.1093/applin/19.3.295.