Bilingualism: Language and Cognition 21 (4), 2018, 826–838 © Cambridge University Press 2017 doi:10.1017/S1366728917000360

Studying texts in a second language: No disadvantage in long-term recognition memory

HELEEN VANDER BEKEN
Ghent University, Belgium
EVY WOUMANS
Ghent University, Belgium
MARC BRYSBAERT
Ghent University, Belgium

(Received: March 08, 2017; final revision received: June 19, 2017; accepted: June 21, 2017; first published online 11 October 2017)

Despite an increase in bilingualism and the use of English as a medium of instruction, little research has been done on bilingual memory for learnt information. In a previous study, we found an L2 recall cost but equal recognition performance in L2 versus L1 when students studied short expository texts (Vander Beken & Brysbaert, 2017). In this paper, we investigate whether there is a recognition cost after a longer delay, which would indicate that the memory trace is weaker in L2. Results showed equal performance in L1 and L2, suggesting that the recall cost is either located at the production level, or that the levels-of-processing effect is mediated by language, with unaffected surface encoding leading to effective MARGINAL KNOWLEDGE *on the one hand, and hampered deep encoding leading to ineffective (uncued) recall. This paper also contains the Dutch vocabulary test we used for native speakers.*

Keywords: bilingualism, learning from text, long-term memory, levels-of-processing effect

Globalisation has led to an increasing number of people that communicate or study in another language than their native tongue. In the European Union, for example, the number of monolinguals has decreased to 46% in 2012 (TNS Opinion & Social). In addition, English is becoming more and more dominant, taking the role of a *lingua franca* (knowledge of some other languages is even decreasing as a consequence; TNS Opinion & Social, 2012). Despite the internationalisation of education and the increasing use of English as a medium of instruction (EMI), little research has been done on the consequences of studying in a second language. Still, with every start of a new academic year, the debate in higher education revives: is it worthwhile to present information, teach, or test students in a language that is not their native one? From an educational perspective, is studying in a second language (L2) a "desirable difficulty" (the perspective that long-term learning occurs through difficulties in learning, e.g., Metcalfe, 2011), a challenge that makes learning just hard enough, or does it obstruct learning possibilities? To answer this question, we need to understand how information is encoded in and retrieved from memory in L2, compared to the first/dominant language (L1).

Declarative memory is traditionally split up between episodic and semantic memory. While semantic memory contains the gist of information about the world, episodic memory contains contextual information tied to the stored event (e.g., Graves & Altarriba, 2014). Information that is processed can be transferred from episodic to semantic long-term memory, in which the contextual information is lost. Neurologically, the hippocampus is responsible for encoding of new – hence, episodic – memories (Hardt, Nader & Nadel, 2013). Within minutes up to hours of this initial hippocampal encoding, neocortical traces are formed. These neocortical traces are the neurological equivalent of semantic memory. In other words: all declarative memory was episodic in its initial stage. Memory consolidation is considered as the reorganization of semantic memory in which hippocampal traces are no longer needed and memory is located in the neocortex only (Hardt et al., 2013). Memory decay can thus be explained by the fact that the hippocampal memory traces are removed during sleep, while the neocortical traces are too weak to remain without the hippocampal connection (though episodic memory remains stored in and retrieved from the hippocampus (Nadel & Hardt, 2011)).

Bilingualism research has explored both types of memory to some extent. The principal theoretical view about bilingual semantic memory has been that meanings of words are stored at a language-independent conceptual level which is connected to all lexicons of a multilingual (e.g., the Revised Hierarchical Model, for a discussion of this model, see Brysbaert & Duyck, 2010). Visual word recognition research has confirmed that, both at the word and sentence level, non-target language knowledge interferes with recognition of a target language (Van Assche, Duyck & Hartsuiker, 2012). This theory accords

* This study was supported by a GOA grant from the Research Council of Ghent University (LEMMA Project).

Address for correspondence:
Heleen Vander Beken, Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, B-9000 Genr, Belgium
heleen.vanderbeken@ugent.be

with an idea found in older memory research (in the 1960–80s, e.g., Alba & Hasher, 1983; Schank, 1972): when people read a text, they do not remember it verbatim but they remember the gist. The so-called *deep structure* of the text remains, though the *surface* form is lost. As a consequence, Schank (1980) concluded, meaning is represented free of language.

Nevertheless, we intuitively expect a disadvantage when reading or studying in L2. Within the language non-selective view, there are three frameworks that would explain L2 disadvantages. The CROSS-LINGUISTIC INTERFERENCE HYPOTHESIS assumes that competition of L1 lexical representations interferes in L2 recognition (see Weber & Cutler, 2004 for auditory recognition). According to some authors (Levy, Mcveigh, Marful & Anderson, 2007; others disagree: Runnqvist & Costa, 2012), this competition results in retrieval-induced forgetting: when you retrieve a concept in one language, this process will hamper retrieval in the other. If L1 representations indeed interfere with L2-recognition, one would expect that reading or studying L2 texts takes more time and that the encoding process is hindered. A second account sets out from the fact that L2 is used less frequently than L1, resulting in weaker linguistic representations (Gollan, Montoya, Cera & Sandoval, 2008). This WEAKER-LINKS HYPOTHESIS directly compares L2-items to low-frequent L1-items. Since familiarity with these items is lower, this account explains that recognition memory for words is better in L2 (the words are less familiar and, as a consequence, more unique in memory; Francis & Gutiérrez, 2012) and entails that L2 semantic representations are less detailed (Finkbeiner, 2002). A third account is located at the level of working memory. The RESOURCE HYPOTHESIS expects that the cognitive load of L2 processing is higher, resulting in less working memory capacity for other processes (Francis & Gutiérrez, 2012; Sandoval, Gollan, Ferreira & Salmon, 2010).

Despite elaborate evidence of language interference in visual word recognition, there are reasons to consider the possibility that memory for text is language-specific. The ENCODING-SPECIFICITY PRINCIPLE (Tulving & Thomson, 1973) states that more information is remembered when the context of encoding and retrieval are similar. Four lines of research provide evidence for this principle. Firstly, in autobiographical episodic memory, people recall more events or more details when they are asked in the language in which the event took place (Marian & Neisser, 2000; Matsumoto & Stanny, 2006; Schrauf & Rubin, 1998). Secondly, in word list recall, more words are recalled in congruent language conditions, even in the weaker language (Nott & Lambert, 1968; Watkins & Peynircioglu, 1983). Thirdly, in listening comprehension, participants are also able to recall more information in the same language condition than in a cross-lingual condition (Marian & Fausey, 2006). Finally, when people read an article in silent or noisy conditions, their recall and recognition performance is better in the context-congruent conditions (Grant, Bredahl, Clay, Ferrie, Groves, Mcdorman & Dark, 1998). If these results translate to other modalities or other types of context, memory for texts might be language-specific, and a lower proficiency may result in lower memory performance. For word list recall and listening comprehension, one might wonder whether these memory traces are part of episodic or semantic memory. Since memory is tested shortly after encoding in these experimental paradigms, the memory consolidation process would not have taken place yet (following Hardt, Nader and Nadel's view, 2013). Hence, encoding specificity is possibly an effect that is limited to episodic or hippocampal memory traces.

Apart from the dichotomy between episodic and semantic memory, there are subtypes within these categories. When discussing semantic memory (or memory in general), we need to take into account what is really tested. People possess large amounts of knowledge, but not necessarily in an active way. Knowledge that cannot be retrieved spontaneously, but can be recognised or retrieved after presentation of a cue, is called "marginal knowledge" (Berger, Hall & Bahrick, 1999; Cantor, Eslick, Marsh, Bjork & Bjork, 2014). A recall test will thus not only test a different type of retrieval, compared to a recognition test (Gillund & Shiffrin, 1984; Haist, Shimamura & Squire, 1992), it will estimate the amount of accessible knowledge, leaving this "marginal knowledge" untouched. In a previous study, we investigated how both recall and recognition for L1 and L2 texts differ (Vander Beken & Brysbaert, 2017). A group of 199 participants studied short expository texts about biology topics within a limited time frame. Afterwards, they received a true/false test about one text and a free recall test about the other. We found no L2 disadvantage in recognition memory, but a significant and rather large disadvantage in L2 recall. These findings indicate that initial encoding was not problematic. Otherwise, there would be a recognition cost as well. However, test performance does suffer from weaker language proficiency in certain conditions. The question is whether this disadvantage is situated merely at the production level, resulting in dissociation between what is known and what can be produced, or at the level of encoding or storage, namely in the richness of the memory trace. Craik and Lockhart's (1972) LEVELS-OF-PROCESSING FRAMEWORK explains that initial encoding processes surface form, while the following stages are responsible for the extraction of meaning. So deeper processing "implies a greater degree of semantic or cognitive analysis" (Craik & Lockhart, 1972, p. 675), also called ELABORATION CODING, which results in a more elaborate and longer lasting memory trace.

ttps://doi.org/10.1017/S1366728917000360

A possible reason to assume a disadvantage in L2 elaboration encoding can be derived from the LANDSCAPE MODEL by van den Broeck, Young, Tzeng, and Linderholm (1999). This theory assumes that a mental model of a text consists of a "landscape" of interrelated concepts (i.e., concepts of biology and text-specific propositions for the texts in our study) that is continuously updated during reading. More specifically, when a concept is activated, it entails cohort activation as well: related concepts are co-activated to a certain extent. Despite the fact that text comprehension suffices for recognising statements about the text, the mental model might be "weaker" in L2. For example, if a domain-specific word is unknown or unfamiliar to the reader, he/she might still understand the sentence or recognise whether a statement is correct, but the concept will not be activated, nor will it activate related concepts. So the semantic richness and activation of the mental model in L2 would be smaller, which is in line with the weaker-links hypothesis that was discussed earlier in this paper. The weaker "landscape"-effect may also be mediated (or enlarged) by lower motivation for reading in L2 (Vander Beken & Brysbaert, 2017), since attention also plays a role in the way concepts are translated to the mental model (van den Broek et al., 1999).

If the mental model is weaker, long term memory would suffer from additional forgetting. Memory traces that are weaker, and less easily recalled, also fade out faster (Craik & Lockhart, 1972). In higher education, information has to be retained for days up to months, and in other real-life situations, retrieval of important information is still relevant after years. Hence, for the current study we decided to test memory for text after longer intervals and employ both an immediate and a delayed recognition test in L1 (Dutch) and L2 (English). The choice for the recognition test was made because (1) there was no difference on the initial recognition test in the previous study, which creates the opportunity to measure additional loss only and (2) the scores on that test were high enough to measure a decrease without dropping to chance level. We do not expect a recognition cost on the immediate test based on the previous study (Vander Beken & Brysbaert, 2017), but there might be a cost for delayed recall. If the recall cost in L2 is due to L2 production only, the rate of forgetting will be similar and none of the language-interference hypotheses will be confirmed, but it could be in line with the encoding-specificity principle. If we find a delayed recall cost, this would suggest that there is a cost at the earlier memory processes, namely encoding (a poorer mental model) or storage, which is in line with the weaker-links hypothesis.

In addition, we test whether memory illusions are more persistent in L2 versus L1. In the context of testing memory, false memory illusions are positive responses to lures in recognition tests (often multiple choice tests). These false memories can be created by merely presenting a false statement repeatedly, increasing the possibility that the statement is later viewed as correct, but it also seems to depend on the performance on the initial test: the illusion is rarely found when the initial answer was correct (Marsh, Roediger, Bjork & Bjork, 2007). Inspired by this finding, we tested whether illusions arose as the consequence of lures in our test. If the memory trace is weaker in L2, due to shallow processing, we expect more illusions to arise in that language.

Method

Participants

A total of 171 first year psychology students from Ghent University participated in partial fulfilment of course requirements and for an additional financial reward. All participants were Dutch native speakers who had studied English in high-school for at least four years and who were regularly exposed to (subtitled) English television programs and English songs. In some of their university courses English handbooks were used, even though teaching took place in Dutch. The data of five students who did not have Dutch as their dominant language were excluded from all analyses. Note that, in this study, L1 was defined in terms of dominant language, not as the first acquired language (though, for most students, the dominant language was also the native language). In addition, seven students were excluded from the analysis because they reported having dyslexia, and another four for other reading or learning disabilities (such as ADD). In the resulting dataset (N = 155), mean age was 19.47 years (sd 4.4); 118 were female students, 33 male (four did not indicate gender). One additional participant was removed from the memory performance analysis for not filling in most of the proficiency tests. Participants were randomly assigned to the conditions.

Materials

Texts

We used two short, English texts from a study of Roediger and Karpicke (2006). Each text covered a topic in the domain of natural sciences: the Sun and sea otters. The English texts were translated into Dutch and the texts were matched between languages on semantics and word frequencies (see Vander Beken & Brysbaert, 2017). The texts were between 248 and 279 words long. They were presented on paper in Times New Roman 10. Line spacing was 1.5 and the first line of every paragraph was indented.

True/false judgement tests

Roediger and Karpicke (2006) divided their texts into 30 ideas or propositions that had to be reproduced. In a previous study (Vander Beken & Brysbaert, 2017), we

used this list as a scoring form for free recall tests and created true/false judgement tests of 46 questions. Thirty true/false questions were derived from the ideas on the free recall scoring form. For example: "The Sun today is a white dwarf star" requires a FALSE response since the text states that "The Sun today is a yellow dwarf star". Next to those literal questions, 10 inferential questions were written: five inferences were based on one proposition, the other five on several propositions in the text. An example of such a question is "The surface of a red giant star is hotter than that of a yellow dwarf star". To respond to that question, the reader has to remember and integrate information about the surface temperature of two of the mentioned star types. In addition, six lure questions were created containing a statement that was not mentioned in the text but was in some way related to a concept in the text. An example of such a statement is "Sea otters live around Alaska", while Alaska was mentioned in the text as the location of an oil spill but not described as sea otters' necessary habitat. All questions were translated to Dutch. The instruction for the test was "Tick the correct answer box for every statement, based on the text you have just read". In the previous study, these questions were checked for passage-dependency in a separate group of participants who did not read the texts, resulting in the exclusion of some questions that were answered better than chance level by this separate group, indicating that they test prior knowledge rather than memory of the texts.

Since it was our goal to test participants' knowledge of the same topic on an immediate and a delayed recall test, we needed two tests for every text. To avoid test effects due to repeated items, we created parallel tests: we selected pairs of questions that were similar in topic and difficulty but did not test the same proposition from the text. For example, when we had one question about the size of sea otters and one about their weight, the first question was included in version A of the test and the second in version B. Difficulty measures were based on test scores on the 46-item version of the test in a previous study (Vander Beken & Brysbaert, 2017). Only the lure questions were repeated in both tests to investigate whether these false propositions led to false memories indeed. These questions were analysed separately and not included in the general analysis. This resulted in two parallel tests of 20 questions for *The Sun* (of which five lure questions) and 18 questions for *Sea Otters* (of which six lure questions).

The tests were administered online, using LimeSurvey (an Open Source PHP web application available through the university). Participants were obliged to answer all questions; answer options were "yes", "no", and "I don't know". The latter option was added to avoid guessing if memory loss is large (this way, chance level scores were avoided).

The texts and the tests can be obtained from the authors for research purposes.

Motivation and text-related questionnaires

After the immediate true/false tests, the participants completed some questions about the texts, concerning prior knowledge, perceived difficulty (of both content and structure), and how interesting the texts were. Next, a questionnaire tapped into their general attitude towards reading and testing. The questionnaire contained single questions for their testing motivation and their self-perceived level of performance relative to fellow students, and several questions about their general reading motivation in Dutch (L1) and English (L2), and their attitudes towards EMI (mostly three questions per sum score). This information can be used to get an insight on how students experience EMI, apart from how they perform. The questionnaires were presented in Dutch to all participants, using 7-point Likert scales.

Subjective assessment of language proficiency

The participants' language background information was assessed with a selection of relevant questions from the Dutch version of the Language Experience and Proficiency Questionnaire (LEAP-Q, Marian, Blumenfeld & Kaushanskaya, 2007; translated by Lisa Vandeberg; adaptation Freya De Keyser, Ghent University, and Marilyn Hall, Northwestern University). This was used to exclude non-dominant speakers of Dutch from all analyses.

Objective L1 proficiency tests

L1 proficiency was measured with a 75-item Dutch vocabulary test in a multiple choice format with four answer alternatives (developed at the department and listed in the Appendix).

Objective L2 proficiency tests

L2 proficiency was measured with the English LexTALE test of vocabulary knowledge for advanced learners of English (Lemhöfer & Broersma, 2012) and Nation and Beglar's (2007) vocabulary size test in multiple choice format. The latter was administered on www.vocabularysize.com, on which researchers can register and set up the test with a log-in code for the participants.

Working memory

Working memory capacity was measured with an automated operation span task programmed in E-prime 2.0.10 (Unsworth, Heitz, Schrock & Engle, 2005).

Text-specific vocabulary knowledge

The delayed English true/false tests were followed by a text-specific vocabulary test in which participants had to

ttps://doi.org/10.1017/S1366728917000360

Table 1. *Mean scores of the interval groups on the various proficiency and intelligence tests (standard deviations between brackets).*

Tests	Day group (N = 49)	Week group (N = 55)	Month group (N = 51)	All (N = 155)
Gender	36F/11M	43F/12M	39F/10M	118F/33M
Age	18.86 (3.49)	19.63 (5.16)	19.92 (4.43)	19.47 (4.44)
Dutch vocabulary (max = 75)	45.80 (7.59)	48.06 (8.53)	46.96 (7.62)	46.99 (7.94)
English LexTALE (max = 100)	72.63 (10.42)	74.95 (11.77)	73.83 (9.24)	73.85 (10.54)
English vocabulary size (max = 140)	95.71 (12.23)	95.75 (14.64)	96 (12.34)	96.09 (13.15)
Operation Span (WM) (max = 75)	58.02 (8.86)	57.18 (13.73)	59.74 (10.50)	58.29 (11.36)

Note: There were missing data points for 8 participants (re-running these comparisons with listwise deletion made no difference). There were no significant differences in the between-groups anovas for the continuous variables. The test statistics can be found at https://osf.io/j8hav/.

explain, translate or give a synonym of the more central or low-frequent words of the texts (10 words for *The Sun*, 14 for *Sea otters)*. Both English and Dutch answers were considered correct.

Procedure

Tests were administered in groups of 33 participants at most. Every participant had to be present for two lab sessions and fill in some questionnaires at home. There were interval groups of 1 day, 7 days or 30 days (plus or minus one day for the two last groups). Students registered online for the sessions, so the interval groups were created based on their availability. We had several subgroups for all three interval groups and selected different times of the day and week to avoid effects of fatigue.

All participants received one text in English and the other in Dutch. The language-text relation was counterbalanced across subjects. All tests were presented in the language of the text. Since there were two parallel tests for every text, the order of these resulted in four conditions (2 text languages x 2 orders of tests) which were counterbalanced across participants. To avoid confusion, all participants first received the text about the Sun, and then the text about sea otters. Combined with the factor interval, the experiment consisted of 12 conditions (2 x 2 x 3 factorial design).

Oral instructions were given in Dutch. At the start of the session, the students were informed that they had to study a text within a limited time frame of seven minutes and that they would be tested afterwards. They were not informed about a delayed recall test in the second session (and at the end of the second session, we checked whether anyone studied the materials during the interval time, which was not the case). They were allowed to highlight sections of the texts or to make some sort of schematic summary, but only on the text itself, which they had to put aside once their study time was up. Testing time was ample with a 4-minute time limit to complete one test, to avoid individual differences in answering time. After the test phase, the procedure (study phase – test phase) was repeated for the second text. In the second session, students filled in the long-term recall tests and carried out the operation span task. All proficiency measures were filled in online via LimeSurvey (unless mentioned otherwise) at home or during the lab time that was left.

Results

Scoring

The true/false judgements were scored dichotomously (correct/incorrect, with "I don't know" as incorrect) with a correction key. The lure questions were analysed separately.

All data are available at https://osf.io/j8hav/ (Open Science Framework).

Testing whether the students were matched in the interval conditions

Because this study tests the effect of interval between-subjects, we first checked whether groups were matched on the control variables we assessed. Table 1 and 2 show that this was the case. There were no significant differences between the three groups if a Dunn-Šidák correction for multiple testing was taken into account ($\alpha = .002846$). The mean L2 proficiency score of 74 for the English LexTALE is comparable to the previous study (M = 72, Vander Beken & Brysbaert, 2017) and is a typical score for this group of Dutch–English participants.

Participants in general had a higher reading motivation in L1 (M = 5.07, SD = 0.81) than in L2 (M = 4.46, SD = 0.89; Wilcoxon signed rank test resulted in V = 8362, $p < .001$), based on a sum score of several questions into reading attitude and motivation. Similar results from a previous study (Vander Beken & Brysbaert, 2017) based on a single question are now confirmed

ttps://doi.org/10.1017/S1366728917000360

Table 2. *Mean scores of the language groups on the self-ratings included in the questionnaire (standard deviations between brackets).*

Self-ratings	Day group (N = 49)	Week group (N = 55)	Month group (N = 51)	All (N = 155)
General motivation				
Test importance (7)	5.04 (1.24)	4.71 (1.36)	4.69 (1319)	4.81 (1.27)
Performance vs. peers (7)	4.02 (0.80)	3.82 (0.86)	3.88 (0.55)	3.90 (0.75)
Dutch academic reading				
Attitude (7)*	4.84 (1.17)	4.68 (1.23)	4.51 (0.90)	4.67 (1.11)
Intrinsic motivation (7)*	5.03 (0.98)	4.87 (0.95)	4.51 (0.90)	4.89 (0.97)
Total motivation (7)*	5.27 (0.77)	4.98 (0.78)	4.96 (0.85)	5.07 (0.81)
English academic reading				
Attitude (7)*	5.77 (0.97)	5.65 (0.97)	5.53 (0.91)	5.65 (0.95)
Intrinsic motivation (7)*	4.50 (1.14)	4.35 (1.12)	4.20 (0.95)	4.34 (1.07)
Total motivation (7)*	4.67 (0.92)	4.4 (0.91)	4.30 (0.81)	4.46 (0.98)
Opinion about use of EMI (7)*	5.68 (0.95)	5.19 (1.30)	5.25 (1.15)	5.36 (1.16)
Dutch language skill				
Reading (10)	9.02 (1.16)	9.2 (0.91)	8.94 (0.84)	9.06 (0.98)
Proficiency (10)*	9.07 (0.87)	9.02 (0.80)	8.84 (0.76)	8.98 (0.81)
English language skill				
Reading (10)	7.67 (1.21)	7.67 (1.16)	7.46 (1.25)	7.60 (1.20)
Proficiency (10)*	7.49 (0.95)	7.37 (1.00)	7.16 (1.07)	7.34 (1.01)

Note: There were no significant differences in the Kruskall-Wallis tests to compare groups. Asterisks indicate sum scores. Likert-scale is indicated between brackets. The test statistics can be found at https://osf.io/j8hav/.

Table 3. *Reliability and correlations of the proficiency and WM measures.*

Tests	Dutch voc. MC	Eng. LexTALE	Vocabulary size	Operation span
Dutch voc. MC	*0.84*	0.44	0.55	0.01
Eng. LexTALE	0.54	*0.77*	0.57	−0.01
Vocabulary size	0.64	0.69	*0.89*	−0.06
Operation span	0.11	−0.01	−0.07	*0.82*

Note: On the diagonal (in italic) is the cronbach's alpha of each test. All numbers above that are original Pearson correlations. The numbers below the diagonal are the correlations corrected for reliability ($r_{xy}/\sqrt{(r_{xx}.r_{yy})}$). One participant did not fill in these four tests (N = 154). There were missing data points for 8 participants, which were omitted by pairwise deletion.

Table 4. *Percentage correct based on the aggregated scores per question.*

	Immediate (all groups)	Day	Week	Month
L1	72.53	65.30	59.40	50.69
L2	71.47	64.68	55.00	46.95

with a more elaborate sum score in this study. The reliability of the objective measures was measured using Cronbach's alpha, which was generally high. Table 3 displays the reliability measures and the correlations between the various measures. There were no motivational measures with M < 4, indicating that our participants were sufficiently motivated to take part in the experiment. When we asked their opinions about the usefulness of EMI at university, we found mildly positive scores as well (see Table 2).

Performance on the memory tests

Memory performance was analysed by means of mixed-effects logistic regression models with the lme4 package (version 1.1-7, Bates, Maechler, Bolker, Walker, Christensen, Singman & Dai, 2014) of R (3.2.2) (R Core Team, 2015). Correctness of the answers was the binary output variable. Language (Dutch vs. English), interval (day/week/month), and session (immediate vs delayed) were included as categorical fixed effects. In a first model, we included the interactions between the three factors (language, interval, and session) and random intercepts and slopes for questions and participants. The R command we used was:

Table 5. *Output of the best fitted glmer-model of the memory scores.*

Fixed effects	Estimate	Std. Error	Z value	P value
(Intercept)	1.21560	0.26067	4.663	3.11e−06***
LanguageEnglish	−0.14871	0.18023	−0.825	0.40932
Intervalmonth	−0.12344	0.18079	−0.683	0.49475
Intervalweek	0.06555	0.17244	0.380	0.70386
Session2	−0.49929	0.15630	−3.194	0.00140***
LanguageEnglish:intervalmonth	0.09214	0.20549	0.448	0.65388
LanguageEnglish:intervalweek	−0.05226	0.20132	−0.260	0.79518
LanguageEnglish:session2	0.13311	0.19942	0.667	0.50447
Intervalmonth:session2	−0.62578	0.20217	−3.095	0.00197**
Intervalweek:session2	−0.40191	0.19842	−2.026	0.04281*
LanguageEnglish:intervalmonth:session2	−0.21029	0.27476	−0.765	0.44404
LanguageEnglish:intervalweek: session2	−0.11609	0.27599	−0.421	0.67402

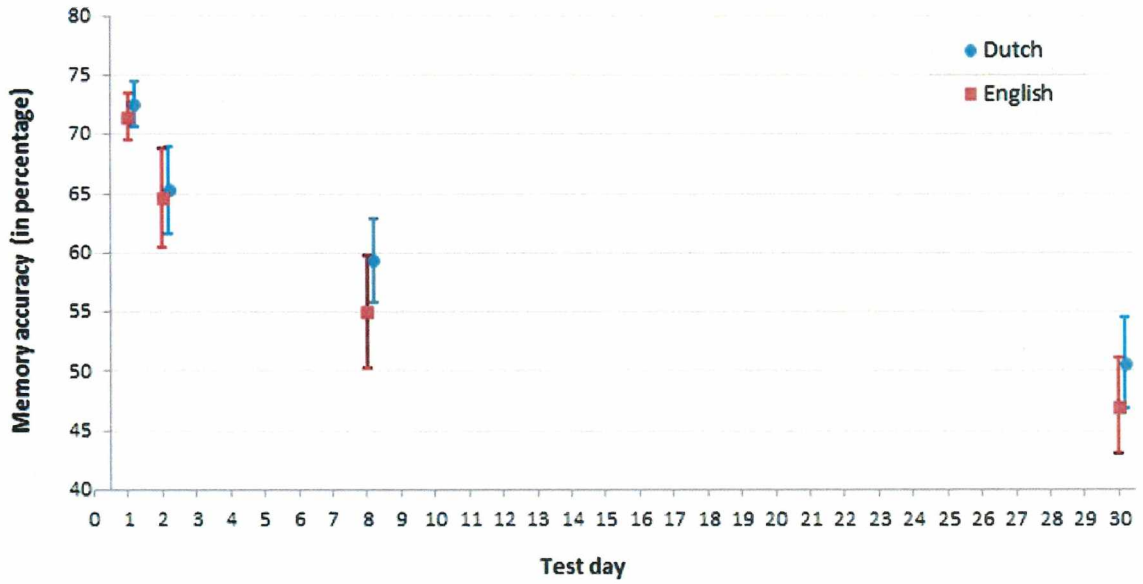

Figure 1. (Colour online) Observed average accuracy over items and participants in L1 (Dutch) and L2 (English). Note that the average on day 1 is aggregated over all interval groups (every participant was tested on day 1) while the other observations are based on separate groups. The SE was multiplied by 1.96 to obtain the confidence interval. More information on the procedure can be found in Brysbaert (2011, pp. 345–355).

$$\text{glmer (correct} \sim \text{language} * \text{interval} * \text{session2} + (\text{language} + \text{interval} + \text{session2}|\text{question}) + (\text{language} + \text{interval} + \text{session2}|\text{id}), \text{mydata}, \text{family} = \text{“binomial”}, \text{verbose} = \text{TRUE})$$

Table 5 displays the output of the analysis. The analysis indicated a significant effect of session ($\beta = -0.50$, SE = 0.16, Z = −3.19, $p < .01$), which means there was a lower chance participants remembered statements correctly in the second session. Furthermore, there were significant interaction effects between session 2 and the week interval ($\beta = -0.40$, SE = 0.20, Z = −2.03, $p < .05$; session 1 and day interval are the reference levels) and between session 2 and the month interval ($\beta = -0.63$, SE = 0.20, Z = −3.10, $p < .01$), indicating that there was more forgetting the longer the interval between tests was. Importantly, language did not have a significant main effect and was not involved in significant interactions. Figure 1 illustrates the rate of forgetting based on the aggregated means in both languages (see Table 4 for the group means).

To check whether the full model hid a small language effect, we ran an extra model excluding the random slopes. This resulted in the same pattern of results. ANOVA comparison showed no significant difference in fit between the models. We also ran two models with only a main

Table 6. *Percentage of false alarms (illusions) based on the aggregated scores per lure question ("yes"-answers). Note: this is based on 5 or 6 questions per test.*

	Immediate (all groups)	Day	Week	Month
L1	8.66	15.77	24.04	40.32
L2	14.52	15.32	34.21	42.62

effect of language (i.e., without the interaction terms involving language) with and without random slopes. In these models there was no effect of language, making us confident that the absence of a significant effect of language is not due to us using a suboptimal model.

Memory illusions

The questions that were added to induce memory illusions were repeated in both sessions and for that reason they were excluded from the general memory performance analysis. If the first test induced memory illusions, more incorrect answers are expected on these questions in the second test. Table 6 displays the percentage of false alarms ("yes"-answers) in all conditions based on 5 or 6 questions per test. These averages clearly show that more false memories arise after a longer interval, despite the option to answer "I don't know". To analyse performance, the data subset was analysed by means of mixed-effects logistic regression models. The binary output variable corresponded to the presence of a false alarm. Language, interval, and session were included as categorical fixed effects. Fitting the full model with all interactions between the three factors (language, interval, and session) and random intercepts and slopes for questions and participants failed due to a convergence error. This is probably due to a lack of variance since most answers were "no"-answers or "I don't know"-answers, both zero values. The data could only be analysed using a glmer-model with the interaction between session and interval, without any other interactions and random slopes. In this model, there was a significant main effect of language (β = 0.44, SE = 0.11, Z = 4.16, p < .001) and session (β = 0.40, SE = 0.19, Z = 2.01, p < .05), so there was a higher chance of false alarms in English compared to Dutch and in the second session compared to the first. There were significant interactions between session and the week interval (β = 1.05, SE = 0.26, Z = 3.97, p < .001) and session and the month interval (β = 1.48, SE = 0.26, Z = 5.73, p < .001), which means the probability of false alarms increases after longer intervals Table 7 displays the output of the analysis.

Discussion

In this experiment, we tested students' recognition memory in Dutch (L1) and English (L2) on an immediate and delayed test, using true/false judgement items from a previous study (Vander Beken & Brysbaert, 2017). Since participants were divided in groups to determine the time of the delayed test, those groups were compared on several objective and subjective measures of proficiency and motivation. There was no difference between groups.

As expected, we did not find an L2 recognition cost on the initial test. Since languages were directly compared in a within-participant design, this robustly confirms the results of the previous study. On the delayed test, there was no significant language effect either. Two conclusions can be drawn from this observation. Firstly, for education, this means that it is no disadvantage for students to be tested on the long term in English, at least for recognition memory. There seems to be no loss of information even though study time was the same in both languages. Secondly, we have found no indication of a disadvantage situated at the level of storage of the mental model and, thus, no evidence for the weaker-links hypothesis.

There are several possible explanations for this finding. It is possible that the recall deficit from the previous study is located at the production level only, which means that people do not remember less in L2 but have more difficulty writing up their recalled memories in L2. Of course, not being able to express the knowledge you have can also

Table 7. *Output of the best fitted glmer-model of the memory illusion scores.*

Fixed effects	Estimate	Std. Error	Z value	P value
(Intercept)	−2.67848	0.34797	−7.698	1.39e−14***
LanguageEnglish	0.43880	0.10539	4.164	3.13*e−05***
Intervalmonth	0.11819²0	0.22569	0.524	0.6005
Intervalweek	−0.08874	0.22472	−0.395	0.6929
Session2	0.38980	0.19371	2.013	0.0442*
Intervalmonth:session2	1.48220	0.25860	5.372	9.94e−09***
Intervalweek:session2	1.04588	0.26333	3.972	7.14e−05***

ttps://doi.org/10.1017/S1366728917000360

be problematic. To confirm this possible explanation, a cross-lingual study (with L2 text – L1 test but also L1 text – L2 test conditions) should show a clear disadvantage of the translation from concept to L2 wordings in all L2 production conditions.

On the other hand, one could argue that a weaker mental model with less "rich" memory traces could still account for unaffected recognition and that memory traces that are weaker produce marginal knowledge. Still, the levels-of-processing framework does suggest that "elaboration coding", i.e., deeper processing with more semantic analysis, results not only in a more elaborate but also a longer lasting memory trace (Craik & Lockhart, 1972). If it were indeed the case and our participants encoded more surface information and less semantic information in the non-dominant language, we would probably have found some long-term memory loss in L2.

These two views seem to exclude one another. Nevertheless, the results from this paper can actually be explained by a combination of opposite effects as well. If the encoding specificity principle can have an effect on studying or reading texts, the unusual context of an L2 study text might create strong contextual cues for retrieving information, compared to L1. In addition, students process less information in L2 than in L1 in the time between the immediate and delayed test, yielding a larger uniqueness of the memory trace in L2 than in L1. So if the encoding of information was less deep in L2 than L1 and the memory trace less strong as a consequence, then there would be a weaker trace in the first instance that suffers less from information interference and is more easily retrieved in a second stage. However, it would be a large coincidence if these two effects were of the exact same size, resulting in a null effect.

Interestingly, Francis and Gutiérrez (2012) showed that the levels-of-processing effect is smaller in L2 than in L1, meaning that shallow processing tasks (e.g., word recognition) yield better recognition performance in the weaker language, but that this advantage decreases for deeper encoding tasks. In other words: deeper encoding tasks mainly improve L1 performance. Taking into account that understanding and remembering a text is a more demanding task in general, this pattern of results is very similar to the combined results from this study and the previous one. The authors explain their observations by a combination of weaker links and resource limitations (Francis & Gutiérrez, 2012).

The experiment also included an attempt to induce memory illusions. More illusions on the delayed test in L2 would suggest that the memory trace is indeed weaker due to shallower processing. We found a main effect of language, indicating that more false memories arise in English. The effects of interval and session, and their interaction, simply suggest forgetting over time. Due to low variance, it was impossible to investigate whether the language effect is mediated over time and whether the memory trace actually fades out more easily. Still, the finding that more lures are remembered as correct in English might be explained by the levels-of-processing effect. If processing is shallower in L2, then maybe a false statement interferes as a new unique memory instead of being rejected based on the contents of the text. This would indicate weaker encoding rather than storage and indirectly strengthens our hypothesis. Nonetheless, despite lower performance on these questions compared to the other questions, we cannot be one hundred percent sure that our attempt to yield false memories with this construction was successful (usually this is tested with multiple choice questions of which one answer is a lure) and we should be careful with strong conclusions about this exploratory element of this study.

To conclude, in this experiment, we found no clear-cut evidence for the weaker-links hypothesis. If L2 memory showed a higher forgetting rate, we would conclude that semantic links are weaker, resulting in weaker memory traces, but this is not the case. Following the logic of Francis and Gutiérrez (2012), the results from this and the previous study could possibly be explained within the levels-of-processing framework: shallow processing tasks on word level result in better L2 recognition than L1 recognition, but there is no such L2 advantage for deeper processing tasks. If you take into account that studying a full text is a more complex task and requires more resources during encoding (according to the resource hypothesis), this effect could translate to our findings. The levels-of-processing effect is larger in a non-dominant language so, perhaps at text level, shallow processing of the L2 texts results in unimpaired long-term recognition performance, but the necessary deeper processing in L2 fails to some extent, compared to L1. (Note that this could still arise within a weaker links framework). Furthermore, we did not find any recognition cost in L2, suggesting that students can be tested in L2 with recognition tests without risking an underestimation of their (possibly marginal) knowledge. In other words, as far as recognition memory goes, the cost-effectiveness of education is not endangered by EMI: the acquired knowledge is retained over a long retention interval (see the introduction section and Berger et al., 1999, p. 438). Further research is necessary to confirm whether the L2 recall cost is actually a production deficit or whether the reason for this disadvantage is more complex and located at the encoding stage of memory. It would also be of great value to explore the current research line with various tests, intervals, and types of bilingual information retention, to discover the commonalities and contrasts between L1 and L2 memory.

Appendix. The 75 items of the Dutch vocabulary test in a multiple choice format with four answer alternatives. The correct answer is underlined.

1. **Successief**: A. Geslaagd, B. Zegevierend, C. Erfelijk, D. Achtereenvolgend
2. **Martelaar**: A. Valsaard, B. Muggenzifter, C. Lijder, D. Prutser
3. **Acteur**: A. Beheerder van goederen, B. Persoon verbonden aan het toneel, C. Ontwerper van auto's, D. Functionaris op treinen
4. **Wauwelen**: A. Dromen, B. Schommelen, C. Spelen, D. Babbelen
5. **Lenigen**: A. Verzachten, B. Leegdrinken, C. Verbuigen, D. Verdedigen
6. **Picaresk**: A. Schilderachtig, B. Met betrekking tot een soldaat, C. Uitbundig, D. Met betrekking tot een schavuit
7. **Bretel**: A. Jas, B. Schoen, C. Broek, D. Pet
8. **Stagnatie**: A. Stilstand, B. Troonsafstand, C. Wisseling, D. Aanpassing
9. **Schrokop**: A. Domoor, B. Schroothoop, C. Vogelschrik, D. Gulzigaard
10. **Knullig**: A. Ontrouw, B. Flauw, C. Onhandig, D. Prullerig
11. **Matig**: A. Krachtig blijvend, B. Voordelig blijvend, C. Efficiënt blijvend, D. Redelijk blijvend
12. **Droedelen**: A. Doelloos tekenen, B. Betekenisloos mompelen, C. Verknoeien, D. Onbewust besmetten
13. **Divan**: A. Tuingereedschap, B. Meubelstuk, C. Auto-onderdeel, D. Operazangeres
14. **Gade**: A. Overtuiging, B. Echtgenoot, C. Burgerwacht, D. Klutser
15. **Dignitaris**: A. Munt van een land, B. Hooggeplaatste ambtenaar, C. Woestijndier, D. Meerderheidsaandeelhouder
16. **Normatief**: A. Opeenhopend, B. Opbouwend, C. Dwingend, D. Mondig
17. **Engerling**: A. Bekrompen man of vrouw, B. Meikever, C. Plant, D. Akelige persoon
18. **Riant**: A. Afwijkend, B. Grappig, C. Verzoeningsgezind, D. Aantrekkelijk
19. **Onbekwaam**: A. Aanstootgevend, B. Niet passend, C. Niet geschikt, D. Niet bezonnen
20. **Paviljoen**: A. Bijgebouw, B. Bijbedoeling, C. Bijfiguur, D. Bijgerecht
21. **Facetoog**: A. Trendy café, B. Insect, C. Nachtdier, D. Donkerblauw oog
22. **Luit**: A. Bouwmateriaal, B. Dier, C. Keukenapparaat, D. Muziekinstrument
23. **Onversaagd**: A. Voortreffelijk, B. Dapper, C. Vrijmoedig, D. Oprecht
24. **Weetal**: A. Oneindig groot getal, B. Betweter, C. Wijze persoon, D. Klein aantal
25. **Patstelling**: A. Positie van waaruit men kan schieten, B. Situatie zonder oplossing, C. Mening die afwijkt, D. Uitspraak van een opschepper
26. **Teint**: A. Specerij, B. Pesterij, C. Kleur, D. Gesp
27. **Voorzaat**: A. Gevelornament, B. Ontkiemend zaad, C. Voorouder, D. Schuine afdekking boven een deur
28. **Slaags**: A. In gevecht, B. Roomsgezind, C. Zich door niets onderscheidend, D. Onderdanig
29. **Kakofonie**: A. Geheimschrift, B. Kabaal, C. Vuile praat, D. Signalisatie
30. **Romig**: A. Slaperig, B. Slordig, C. Dik en vloeibaar, D. Met lijm bedekt
31. **Schimpen**: A. Scheuren, B. Schelden, C. Schudden, D. Schuiven
32. **Rups**: A. Hondjes, B. Larve, C. Taartjes, D. Aardig
33. **Opsmuk**: A. Opschudding, B. Versiering, C. Beveiliging, D. Ontplooiing
34. **Laakbaar**: A. Niet te vertrouwen, B. Afkeurenswaard, C. Afschuwwekkend, D. Aan lijden onderhevig
35. **Woelig**: A. Tactvol, B. Turbulent, C. Delicaat, D. Ontroerd
36. **Verguld**: A. Als gunst toegestaan, B. Met smaad bejegend, C. Als voedzaam verkocht, D. Met goud bedekt
37. **Publiekelijk**: A. Bevallig, B. Aansprakelijk, C. Kostbaar, D. Openbaar
38. **Exploitatie**: A. Een niet-democratische staatsvorm, B. Opgeblazenheid, C. Gebruik maken van, D. Loslaten van een orgaan
39. **Masochist**: A. Iemand die graag anderen pijn doet, B. Iemand die geen gezag erkent, C. Iemand die gemakkelijk van mening verandert, D. Iemand die graag vernederd wordt
40. **Ontredderd**: A. In veiligheid, B. Troosteloos, C. Vertederd, D. In gevaar
41. **Relaas**: A. Verslag, B. Troost, C. Steun, D. Familielid

ttps://doi.org/10.1017/S1366728917000360

42. **Macaber**: A. Griezelig, B. Kleurrijk, C. Ambitieus, D. Onbetrouwbaar
43. **Grimeren**: A. Beschadigen, B. Beschilderen, C. Beschermen, D. Beschuldigen
44. **Hekelen**: A. Overgieten, B. Spelen, C. Inzouten, D. Bekritiseren
45. **Platvloers**: A. Languit, B. Vlak, C. Grof, D. Effen
46. **Gong**: A. Slaginstrument, B. Sleepinstrument, C. Blaasinstrument, D. Houtinstrument
47. **Perikelen**: A. Rondkijken, B. Slachten, C. Moeilijkheden, D. Aanmoedigen
48. **Rekruut**: A. Soldaat, B. Reglement, C. Onmens, D. Hoedanigheid
49. **Exorcisme**: A. Het misbruiken van vertrouwen, B. Het vernielen van cultuurgoederen, C. Het onderdrukken van emoties, D. Het uitdrijven van duivels
50. **Xenofoob**: A. Waterafdrijvend, B. Vreemdelingenhater, C. Iemand met pleinvrees, D. Muziekinstrument
51. **Finesse**: A. Lenigheid, B. Lichaamsconditie, C. Bijzonderheid, D. Levendigheid
52. **Tequila**: A. Schelp, B. Pannenkoekje, C. Monster, D. Alcohol
53. **Verbolgen**: A. Taboe, B. Beduusd, C. Verbluft, D. Boos
54. **Tendens**: A. Aantrekkelijkheid, B. Neiging, C. Verleiding, D. Bekoring
55. **Prieel**: A. Uit overtuiging, B. Tuinhuis, C. Oorspronkelijk, D. Gedeeltelijk
56. **Betichten**: A. Aanvechten, B. Betreuren, C. Bedriegen, D. Aanklagen
57. **Nerf**: A. Marterachtige, B. Bladader, C. Zenuwlijder, D. Sukkel
58. **Guitig**: A. Voordelig, B. Bevorderlijk, C. Plechtig, D. Speels
59. **Stramien**: A. Geheim, B. Moeizaam, C. Patroon, D. Zeer hoog
60. **Wrok**: A. Bouwval, B. Keukengerei, C. Haat, D. Gierigaard
61. **Courant**: A. Vloeiend, B. Gebruikelijk, C. Toegeeflijk, D. Te voet
62. **Castagnetten**: A. Fruit, B. Kleren, C. Muziek, D. Groenten
63. **Verijdelen**: A. Onderdrukken, B. Onderwerpen, C. Onderzoeken, D. Onderbreken
64. **Heling**: A. Aanraken van heilige voorwerpen, B. Aannemen van gestolen goed, C. Aanmanen tot actie, D. Aandrijven van voertuigen
65. **Seniel**: A. Breekbaar, B. Zwakzinnig, C. Verplaatsbaar, D. Onvast
66. **Vergen**: A. Keren, B. Ontdoen, C. Reinigen, D. Eisen
67. **Drek**: A. Vocht, B. Lucht, C. Bloed, D. Mest
68. **Lijvig**: A. Saai, B. Dik, C. Opwindend, D. Lichamelijk
69. **Zeis**: A. Graaien, B. Maaien, C. Naaien, D. Zaaien
70. **Rekwisieten**: A. Beperkingen, B. Benodigdheden, C. Afbakeningen, D. Versnaperingen
71. **Dorpel**: A. Onverstaanbare spraak, B. Kleine hond, C. Kleine stad, D. Deur
72. **Inham**: A. Weiland, B. Nageboorte van een merrie, C. Baai, D. Achterbout van een varken
73. **Overstelpen**: A. Overwerken, B. Overhalen, C. Overladen, D. Overtreden
74. **Feeks**: A. Schroevendraaier, B. Boor, C. Tang, D. Hamer
75. **Dressoir**: A. Werktuig, B. Boom, C. Klimaat, D. Meubelstuk

References

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, *93*(*2*), 203–231. https://doi.org/10.1037/0033-2909.93.2.203

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singman, H., & Dai, B. (2014). Package "lme4": Linear Mixed-Effects Models using "Eigen" and S4.

Berger, S. A., Hall, L. K., & Bahrick, H. P. (1999). Stabilizing access to marginal and submarginal knowledge. *Journal of Experimental Psychology*, *5*(*4*), 438–447.

Brysbaert, M. (2011). *Basic statistics for psychologists*. Palgrave Macmillan.

Brysbaert, M., & Duyck, W. (2010). Is it time to leave behind the Revised Hierarchical Model of bilingual language processing after fifteen years of service? *Bilingualism: Language and Cognition*, *13*(*3*), 359–371. https://doi.org/10.1017/S1366728909990344

Cantor, A. D., Eslick, A. N., Marsh, E. J., Bjork, R. a., & Bjork, E. L. (2014). Multiple-choice tests stabilize access to marginal knowledge. *Memory & Cognition*, *43*, 193–205. https://doi.org/10.3758/s13421-014-0462-6

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of Processing: A Framework for Memory Research. *Journal of Verbal Learning and Verbal Behavior*, *11*(*6*), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Finkbeiner, M. S. (2002). *Bilingual lexical memory: Towards a psycholinguistic model of adult l2 lexical acquisition,*

representation, and processing (Doctoral dissertation): The University of Arizona.

Francis, W. S., & Gutiérrez, M. (2012). Bilingual recognition memory: Stronger performance but weaker levels-of-processing effects in the less fluent language. *Memory & Cognition*, *40*(3), 496–503. https://doi.org/10.3758/s13421-011-0163-3

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *9*, 1–67.

Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language*, *58*(3), 787–814. https://doi.org/10.1016/j. jml.2007.07.001

Grant, H. M., Bredahl, L. C., Clay, J., Ferrie, J., Groves, J. E., Mcdorman, T. A., & Dark, V. J. (1998). Context-Dependent Memory for Meaningful Material: Information for Students. *Applied Cognitive Psychology*, *12*, 617–623.

Graves, D. F., & Altarriba, J. (2014). False Memories in Bilingual Speakers. In R. R. Heredia & J. Altarriba (Eds.): Foundations of Bilingual Memory. Springer. 205–221.

Haist, F., Shimamura, A. P., & Squire, L. R. (1992). On the Relationship Between Recall and Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(4), 691–702.

Hardt, O., Nader, K., & Nadel, L. (2013). Decay happens: The role of active forgetting in memory. *Trends in Cognitive Sciences*, *17*(3), 111–120. https://doi.org/10.1016/j.tics.2013.01.001

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

Levy, B. J., Mcveigh, N. D., Marful, A., & Anderson, M. C. (2007). Inhibiting Your Native Language: The Role of Retrieval-Induced Forgetting During Second-Language Acquisition. *Psychological Science*, *18*(1), 29–34. https://doi.org/10.1111/j.1467-9280.2007.01844.x

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing Language Profiles in Bilinguals and Multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*(August), 940–967. https://doi.org/10.1044/1092-4388(2007/067

Marian, V., & Fausey, C. M. (2006). Language-dependent memory in bilingual learning. *Applied Cognitive Psychology*, *20*, 1025–1047. https://doi.org/10.1002/acp.1242

Marian, V., & Neisser, U. (2000). Language-dependent recall of autobiographical memories. *Journal of Experimental Psychology: General*, *129*(3), 361–368. https://doi.org/10.1037// 0096–3445.129.3.361

Marsh, E. J., Roediger, H. L. I., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, *14*(2), 194–199. https://doi.org/10.3758/BF03194051

Matsumoto, A., & Stanny, C. (2006). Language-dependent access to autobiographical memory in Japanese-English bilinguals and US monolinguals. *Memory (Hove, England)*, *14*(3), 378–390. https://doi.org/10.1080/09658210500365763

Metcalfe, J. (2011). Desirable Dificulties and Studying in the Region of Proximal Learning. In A. S. Benjamin (Ed.), *Successful Remembering and Successful Forgetting: A Festschrift in Honor of Robert A. Bjork* (pp. 1–27). Psychology Press.

Nadel, L., & Hardt, O. (2011). Update on Memory Systems and Processes. *Neuropsychopharmacology*, *36*(1), 251–273. https://doi.org/10.1038/npp.2010.169

Nation, P., & Beglar, D. (2007). A vocabulary size test. *JALT*, *31*(7), 9–12.

Nott, C. R., & Lambert, W. E. (1968). Free Recall of Bilinguals. *Journal of Verbal Learning and Verbal Behavior*, (7), 1065–1071.

R Core Team (2015). R: A language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. Vienna, Austria. https://www.R-project.org.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–55. http://doi.org/10.1111/j.1467-9280.2006.01693.x

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

Sandoval, T. C., Gollan, T. H., Ferreira, V. S., & Salmon, D. P. (2010). What causes the bilingual disadvantage in verbal fluency? The dual-task analogy. *Bilingualism: Language and Cognition*, *13*(132), 231–252. https://doi.org/10.1017/S1366728909990514

Schank, R. C. (1972). Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, *3*(4), 552–631. https://doi.org/10.1016/0010-0285(72)90022-9

Schank, R. C. (1980). Language and memory. *Cognitive Science*, *4*, 243–284. https://doi.org/10.1016/S0364-0213(80)80004-8

Schrauf, R. W., & Rubin, D. C. (1998). Bilingual Autobiographical Memory in Older Adult Immigrants: A Test of Cognitive Explanations of the Reminiscence Bump and the Linguistic Encoding of Memories. *Journal of Memory and Language*, *39*, 437–457.

TNS Opinion & Social (2012). *Europeans and their Languages*. *Special Eurobarometer 386*. Brussels: European Commissean. ec.europa.eu/public_opinion/archives/ebs/ebs_386_en.pdf

Tulving, E., & Thomson, D. M. (1973). Encoding Specificity and Retrieval Processes in Episodic Memory. *Psychological Review*, *80*(5), 352–373.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505.

Van Assche, E., Duyck, W., & Hartsuiker, R. J. (2012). Bilingual word recognition in a sentence context. *Frontiers in Psychology*, *3*(June), 174. https://doi.org/10.3389/fpsyg.2012.00174

van den Broek, P., Young, M., Tzeng, Y., & Linderholm, T. (1999). The Landscape Model of Reading: Inferences and the Online Construction of Memory Representation.

In H. van Oostendorp & S. R. Goldman (Eds.), *The Construction of Mental Representations During Reading* (p. 404). Lawrence Erlbaum Associates.

Vander Beken, H., & Brysbaert, M. (2017). Studying texts in a second language: The importance of test type. *Bilingualism: Language and Cognition* 1–13. https://doi.org/10.1017/S1366728917000189

Watkins, M. J., & Peynircioglu, Z. F. (1983). On the Nature of Word Recall: Evidence for Linguistic Specificity. *Journal of Verbal Learning and Verbal Behavior*, *22*, 385–394.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50*(*1*), 1–25. https://doi.org/10.1016/S0749-596X(03)00105-0