ON REPLICATION AND LEXICAL FREQUENCY IN PSYCHOLINGUISTICS

Heather Swanson, Bethany MacLeod, and Daniel Siddiqi
Carleton University

1. Introduction

As the field of linguistics increasingly incorporates the experimental methods of psychology, it is becoming more and more important for us to recognise that along with the empirical strengths we gain through these methods, we also inherit the methods' weaknesses. In particular, we inherit the need to replicate experimental findings (Diener & Biswas-Diener 2016), a problem the traditional methods of theoretical linguistics have not necessarily had to grapple with. This paper marks the first of an ongoing research program at Carleton University dedicated to the replication of published findings of psycholinguistic experiments in order to fulfill a basic mandate of the scientific method. The results of our first replication confirm the need for replication in psycholinguistics in rather unforeseen ways and suggest more serious problems with psycholinguistic experimental design, especially with regard to experiments that test or control for lexical frequency. The purpose of this paper is to explain our findings and make a case for the importance of replication in linguistic studies.

The study we chose to replicate was McCormick, Brysbaert, and Rastle (2009) (henceforth MBR), which investigates whether lexical frequency plays a role in the lexical decomposition of words during word recognition. We chose to replicate this study because lexical decomposition is a frequent topic of study at Carleton and because MBR's methods were straightforward for us to reproduce in our lab. Furthermore, the outcome of studies like MBR are important because determining if frequency affects decomposition has consequences for our choice of lexical processing model, such as single or dual route models (Taft 1994; Baayen, Dijkstra, & Schreuder 1997; Kuperman, Bertram & Baayen 2008). We had no special concerns with the methodology of MBR, nor did we have any a priori reason to suspect our findings would be any different from MBR.

What is meant by "replication" is not agreed upon in the field of experimental psychology, and by extension psycholinguistics (Porte 2013). One definition assumes that the purpose of replication is to run an experiment that is identical to the original, including the same materials, procedure, and data analysis, differing only by the specific individuals running and participating in the study. Under this approach, a replication is simply a duplication of the original and if the findings of the replication are the same as the original then we can say that they have been replicated. A stronger definition assumes that the purpose of replication is to determine whether the answer to the research question that the original study found is, in fact, true. Under this approach, we must do more than duplicate the original study; we must create a new experiment that aims to answer the

Actes du congrès annuel de l'Association canadienne de linguistique 2017.
Proceedings of the 2017 annual conference of the Canadian Linguistic Association.

same research question and then compare the findings of the two experiments. If the original finding was not spurious, then we should be able to expect that we will find the same result even if we introduce some differences in the methods of the replication, including different participant population, specific materials, exact procedure, and choice of data analysis techniques[1].

In this paper we will refer to the former type of replication (where the experiments differ only in subjects and experimenters) as a "reproduction" and the latter type (where new materials or procedures could be included) as a "replication". We performed both a reproduction of MBR, which aims to change as little of the original as possible, and a replication, which aims to confirm the methodology and materials as well as the findings. In our replication, many of the details were the same as MBR; however, different participants, a different corpus, different controls, and a different statistical model were used to verify each stage of the experimental process. We proxied a reproduction study (since we didn't have all the original materials) by using the same corpus as MBR.

Our results for both studies proved surprising. The results of the replication study (which used a different corpus to determine lexical frequency from MBR) confirm the results of MBR, where no effect of relative lexical frequency was found. However, there was a significant mismatch between the lexical frequencies used in MBR and those used in our replication study. These lexical frequencies determined how the target stimuli were sorted. The difference in frequencies between MBR and our replication caused the target stimuli to be sorted so differently that a meaningful comparison of the two is impossible. This means that the result of the replication endeavor was the finding of a potentially serious design flaw: the assumption of a particular corpus' frequency data. This potential design flaw has very serious consequences because of the prevalence of using frequency data in psycholinguistic experiments without controlling for corpus variation (if that is even possible).

On the other hand, the reproduction study (which used the same corpus as MBR) found an effect of lexical frequency on decomposition where MBR did not. We thus have contradictory results of our replication project. Our replication study confirmed the findings of MBR, but identified a serious design problem inherent in controlling for or studying lexical frequency. Our reproduction study refuted MBR and found opposite results. Because we now have three studies that say three different things, at this point we assume that we should not be making any serious theoretical claims about lexical processing. However, what we have found has serious methodological ramifications for psycholinguistics.

This paper will be structured as follows. Section 1 will briefly summarize MBR, including why the study was done. Section 2 will detail the replication study. Section 3

[1] Of course, the way in which we make these methodological changes is crucially important. Some methodological differences will cause the experiment to be asking a different question. In our definition of replication, some of the details of the experimental design may differ from the original, but not in such a way or to an extent that we would expect them to change the nature of the findings. The idea is to determine whether the original finding is robust to minor differences in experimental design, not major changes.

will detail the reproduction study. Section 4 will attempt to make some sense of our findings. Section 5 discusses the findings in light of the project of replication.

2. McCormick et al. (2009)

Prior research on theories of visual word recognition shows evidence of lexical decomposition in morphologically-complex words. Routine lexical decomposition is supported by both single-route models and parallel dual-route models. Under a single route model, morphologically-complex words are always decomposed into their component morphemes during recognition (e.g. Taft 1994). Parallel dual route models tend to come in two varieties. In one variety, morphologically-complex words are recognized through the combination of decomposition and direct lexical retrieval (Baayen, Dijkstra, & Schreuder 1997; Kuperman, Bertram, & Baayen 2008). Other varieties use the horse race metaphor where a word is recognized either by lexical decomposition or by direct lexical retrieval, depending on which is the fastest route. Previous studies have suggested that words with lower lexical frequency are more likely to be decomposed while words with higher frequency are more likely to be stored and accessed whole. Most of these studies, however, have only included words that are relatively low in frequency (Longtin, Segui & Halle 2003; Rastle, Davis, & New 2004).

The goal of MBR was to include both low and high frequency words in what they considered to be a better test of the effect of lexical frequency on decomposition in word recognition. The sixty native English speakers who took part in the study performed a lexical decision task. They were presented with strings of letters on a computer screen and their task was to decide as quickly and as accurately as possible if the string was a real word or not. The participants were not aware that a masked prime had been presented before the string for 42ms. Response accuracy and response time (RT) were recorded.

The stimuli included 120 prime-target pairs. The primes were morphologically-complex words and the target words were the stem of the prime. For example, the prime *national* was paired with the target *nation.* Forty of the primes were words that were at least two times more frequent than the stem (the high frequency condition, such as comfortable-COMFORT[2]) and another forty were at least two times less frequent than the stem (the low frequency condition, such as fragility-FRAGILE). Word frequencies were determined using CELEX (Baayen, Piepenbrock, & van Rijn 1993). The remaining forty primes were pseudoderived words (the pseudoword condition, including forms such as *purplity*, whose stem would be PURPLE). The pseudowords have no lexical frequency, but since they are pseudoderived they could still be decomposed during word recognition. The study also included control primes that were unrelated to the target words as well as non-word targets.

If the primes are decomposed, then we would expect a decrease in RT when the prime is related to the target. However, the dual-route models discussed above predict

[2] The prime is given in lowercase and the target word in uppercase letters.

that this priming effect would only be found when the prime is low frequency since, under the horse-race style of model, high frequency words are retrieved whole and not decomposed during recognition. In MBR's study then, if the priming effect is only found for the low frequency and pseudoword conditions, this would support the dual-route models (such as Portin et al. 2008 and Hay 2001). However, if the priming effect is found for the high frequency condition as well, this would suggest that lexical decomposition occurs regardless of the frequency of the prime, providing support for the single-route or parallel dual-route models (such as Baayen, Dijkstra, & Schreuder 1997). The results of MBR found no statistically significant difference in priming effect by condition, suggesting that lexical frequency has no effect on decomposition and providing support for single-route or parallel dual-route models of lexical access.

3. Replication Study

In our replication study, we aimed to recreate the entire design of MBR, verifying each part of the design. We tested 59 undergraduate volunteers from Carleton University. All participants were native English speakers, 18 years of age or older. We included the same 120 prime-target pairs as MBR; however, since the filler words and the non-word targets were not provided by MBR, we had to generate our own. As MBR did, we included forty pairs of unrelated primes and targets as fillers. Following the methodology of MBR, in all forty pairs, the target words were real words, two thirds of the unrelated primes were real suffixed words, and one third were pseudoderived words. These filler targets were matched to the 120 experimental targets in terms of frequency, length, and neighbourhood size. We also generated a total of 160 non-word targets, derived from the 120 target words and forty real word fillers, which were selected for the NO response of the lexical decision task. These non-words were matched to the real target words for length and neighbourhood size. In the original study, the non-word targets were matched groupwise to the experimental primes on morphological status, lexical status, length, and frequency. In this study, we used a random selection of the same real word primes used for the real-word targets to precede the non-word targets.

MBR used CELEX (Baayen, Piepenbrock, & van Rijn 1993) to determine lexical frequencies of all the stimuli. CELEX is a large corpus, comprising at least 17 million words and is a mix of both British and American English. Since the participants in MBR were students from Royal Holloway, University of London, using a corpus based on British English makes sense. Since the participants in our replication were Canadian, we chose to use a corpus based on North American English to determine lexical frequencies. We used the Corpus of Contemporary American English (Davies 2008), which is also a very large corpus, comprising more than 520 million words, to obtain the frequencies of the stimuli and then followed MBR's approach to grouping the stimuli. That is, primes fell into the high frequency condition if they were at least twice as frequent as their stem (the target) and fell into the low frequency condition if they were least two times less frequent than their stem.

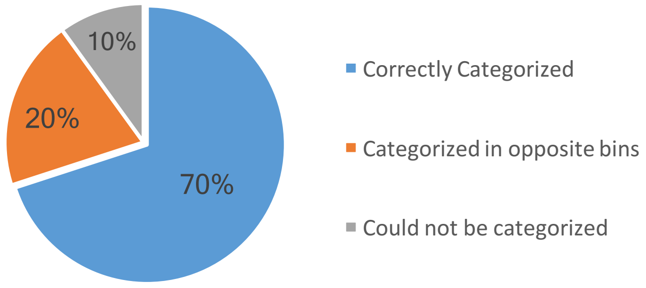
Following MBR, all the target conditions were divided into two lists so that half of the targets in each list would be preceded by a related prime and the other half by an unrelated prime. Each participant saw only one of these lists and made 320 lexical decisions.

3.1 Differences in lexical frequency between the corpora

Since COCA and CELEX are two different corpora, we would not expect the raw frequency values to be identical. However, we might expect that the relative frequencies of the members of each prime-target pair would be similar. That is, if a particular prime were more than twice as frequent as its stem in COCA, we would expect that it would be more than twice as frequent as its stem in CELEX. Perhaps the ratio would not be identical, and we could expect some variation due to the dialectal differences between the two corpora, but we would also expect that the majority of the prime-target pairs would be binned in the same way using the two corpora.

However, when we compared the binning of the prime-target pairs as performed using the COCA frequencies to that using the CELEX frequencies, we found that a substantial portion of the pairs were binned differently. The frequencies drawn from the two corpora were different enough that we had to rearrange which targets were in which condition 30% of the time.

Figure 1. Breakdown of target pairs that were rearranged following COCA frequencies



As shown in Figure 1, 20% of the prime-target pairs were put in opposite bins according to the COCA frequencies (i.e. some prime-target pairs in the high frequency condition according to CELEX were found to have a prime less frequent than the stem using COCA frequencies or, in the case of the low frequency primes, a prime was found to be more frequent than the stem according to COCA when it was no more than half as frequent according to CELEX). 10% failed to meet the criteria to be put into either bin (i.e. they could not be grouped according to COCA frequencies because the prime was neither two times more frequent than its stem, nor was it no more than half as frequent). These discrepancies are illustrated in Figures 2 and 3. Figure 2 shows a sample binning of 40 target-prime pairs using CELEX frequencies: 20 pairs are categorized in the high frequency condition and the other 20 in the low frequency condition. When we categorize these same 40 pairs using COCA frequencies, however, 20% of them end up in the opposite bins. These are given in red in Figure 3. An additional 10% do not fall into either bin. These are given in purple in Figure 3.

Figure 2. Sample of binning using CELEX frequencies

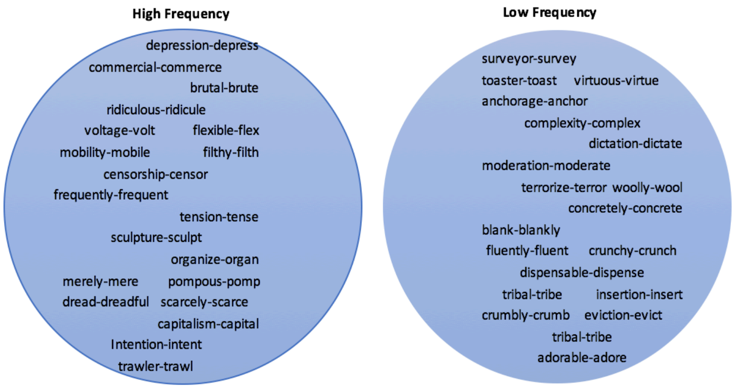
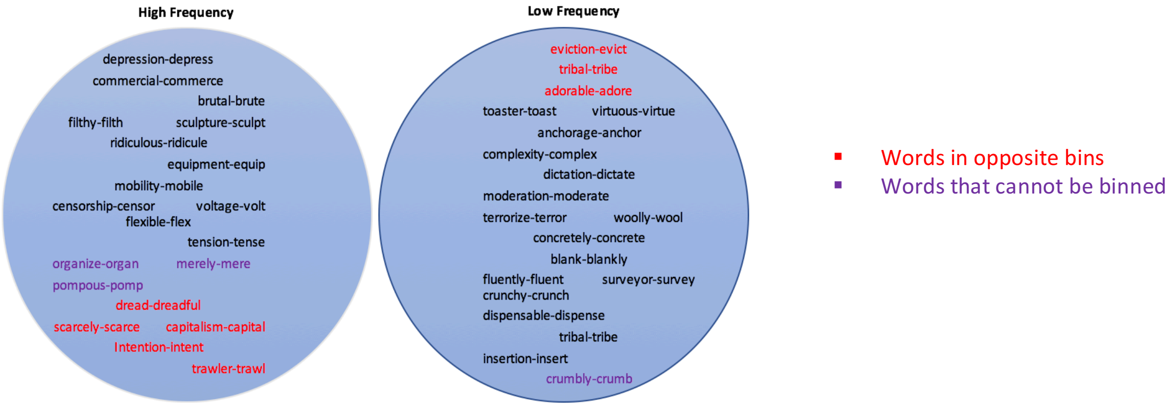


Figure 3. Sample of binning using COCA frequencies



3.2 Procedure

This experiment was run using PsychoPy (Pierce 2007) as opposed to the DMDX software used by MBR and took place in a quiet room on the Carleton University campus. Prior to beginning the experiment, participants completed a linguistic background questionnaire. Participants were then seated in front of a computer and were asked to respond to a string of letters that appeared on the screen. Their task was to decide whether the string of letters they saw was a real English word or not. They were instructed to respond quickly but to not rush through the responses. A forward mask appeared for 500ms, followed by either a related or unrelated prime for 42ms. The target word was then presented and stayed on the screen until a decision was made, using two different keys on a keyboard. Response time and response accuracy for each decision was recorded using PsychoPy.

3.3 Results

As MBR did, we discarded RT for incorrect responses and outlying data points (RT over 2500ms and under 150ms). Table 1 below shows the mean RT for each condition for the related and control primes as well as the priming effect (the difference between the mean RT for the control primes and the related primes).

Table 1. Mean of response times by condition using COCA frequencies

Condition	High frequency	Low frequency	Pseudoword
Related prime	82.92	82.73	83.74
Control prime	85.23	82.82	85.59
Priming effect	2.31	0.09	1.85

Unlike MBR, who used a three-factor ANOVA, we analyzed the data using a mixed effects model in R (R Core Team 2016). Data that could not be binned according to the COCA frequencies were eliminated. The dependent variable was RT. Fixed effects included relatedness of the prime (reference level is 'unrelated'), frequency (3 levels: low, high, pseudo), and an interaction between relatedness and frequency. Participant, target, and prime were random effects. None of the main effects or interactions were found to be significant. No main effects and no other interactions were significant. As discussed above, MBR found that there was no effect of lexical frequency on response time. We replicated those findings. There was no statistically significant difference found between the response times for any of the conditions.

4.0 Reproduction Study

The discovery that the COCA and CELEX frequencies mismatched to the extent that they did mandated that we analyze our data a second time using the CELEX frequencies. In order to proxy a more faithful replication, which we call here a reproduction, we ran our statistics again except we binned our targets using the CELEX frequencies as MBR did. Because we used the CELEX statistics, the 10% of targets that were previously rejected as unbinnable were kept. Otherwise, everything between the replication and the reproduction is kept the same.

4.1 Results

The data were again analyzed using a mixed effects model in R (R Core Team, 2016). The dependent variable was RT. Fixed effects included relatedness of the prime (reference level is 'unrelated'), frequency (3 levels: low, high, pseudo), and an interaction between relatedness and frequency. Participant, target, and prime were random effects. The interaction term where low frequency was compared to high frequency was significant ($t = -2.21$, $p < 0.05$), with a negative coefficient ($\beta = -0.042$) indicating that the RT was statistically significantly shorter when a related prime was high frequency as compared to when it was low frequency. No main effects and no other interactions were significant.

These findings are different from those of MBR, who found that there was no statistically significant difference in RT between the conditions. This finding, that there is a greater priming effect when the prime is high frequency than when it is low frequency, is very surprising considering the existing models of lexical access. As discussed earlier, the single-route models and parallel dual-route models would predict no difference in priming effect between the frequency conditions and the horse-race style dual-route models would predict a greater priming effect for low frequency primes than for high. Additionally, the very small magnitude of the priming effect for all three conditions (as shown in Table 2 below) is also very surprising; although the difference between the priming effect in the high and low frequency conditions was statistically significant, both effects are very low and close to 0ms, so it may be that this difference has no practical significance. However, from a statistical perspective, we have found two different results depending on which corpus we used: when the COCA frequencies are used, there was no statistically significant difference in priming effect depending on frequency, but when the CELEX frequencies were used, there was a greater priming effect found in high frequency words.

Table 2. Mean of response times by condition using CELEX frequencies

Condition	**High frequency**	**Low frequency**	**Pseudoword**
Related prime	82.5	83.94	83.74
Control prime	85.76	82.35	85.59
Priming effect	3.26	-1.59	1.85

5.0 Discussion

There are several threads that we need to discuss here to interpret our findings, which we think are rather unexpected. First, we discuss the importance of choice of corpora in psycholinguistic experiment design. The most important finding of our study is how important this choice seems to be. Second, we will reflect on the importance of replication in contemporary linguistic research. Third, we will attempt to interpret our results (which are somewhat baffling).

5.1 Choice of corpora

In determining the lexical frequencies of their target words, MBR used CELEX (Baayen, Piepenbrock, & van Rijn 1993), as their corpus. CELEX is a large corpus, comprising at least 17 million words and is a mix of both British and American English. We used COCA (Davies 2008) in the replication to verify the lexical frequencies and the groupings of the original study. COCA is also a very large corpus, comprising more than 520 million words. However, it only uses American English, which arguably better represents our subjects. Both corpora are frequently used in the literature, and both are well regarded.

Recall that the main target materials in MBR were divided according to relative high and relative low frequency. The high frequency condition included primes that were at least two times more frequent than its stem word, while the low frequency condition included primes that were no more than half as frequent as its stem word. Seventy percent of the words were binned the same way in both corpora. However, thirty percent of the word frequencies did not match between the databases. Twenty percent of the lexical items that were binned a certain way according to CELEX were put in opposite bins according to COCA. The last ten percent were words that could not be grouped according to COCA because the derived words were neither at least two times more frequent than their stems, nor were they at most half as frequent.

The discrepancies between the two corpora show the ramifications that the choice of corpora can have on a psycholinguistic experiment. It is fair to say that our COCA replication of MBR is not really the same experiment. Any answers it found would be answers to different questions because the words were sorted differently. Importantly, though, MBR and our COCA replication think they are asking the same questions and have found the same answers even though the realities they are putatively testing are mutually incompatible**.** This is a fundamental flaw in our replication that reveals a fundamental flaw in the original MBR design: MBR assumed that the CELEX frequencies were an accurate representation of reality and could thus be tested. But this assumption was clearly false because COCA did not have the same representation of reality. Further, our COCA replication also makes the same fundamental error by assuming that COCA fairly represents reality. It is simply true, as is well known in the corpus linguistics literature (van Heuven, et al. 2014; Burgess & Livesay 1998; Erker & Guy 2012), that it is just not safe to assume that a particular corpus is an accurate representation of lexical frequency. Corpora are different and make different design choices that can lead to different results.

This type of mismatch has a profound butterfly effect on psycholinguistics. The experiments here happen to be studying lexical frequency, but controlling for lexical frequency even when studying other effects is pervasive in the field of psycholinguistics. We sampled twelve studies from the literature, all of which relied heavily on lexical frequency. Specifically, we reviewed whether those studies used more than one corpus in determining frequency. These studies ranged from 1985 to 2013, and studied English, French, or Spanish. We found that all twelve of the studies used only one corpus. Six of the studies used CELEX as the primary and only corpus, such as Merkx, et al. (2011).

While CELEX is a relatively large corpus, other databases were not. For example, in Perea et al. (2005), the Dictionary of Word Frequencies in Spanish, containing two million words, was used (Alameda & Cuetos 1995). In another study by Majerus & Van der Linden (2003), the BRULEX database was used (Content et al. 1990). We glean from this survey that as a generalization, careful selection of large corpora does not seem to occur in many studies that investigate lexical frequencies (let alone those that just control for it). Additionally, these studies almost never use or reference more than one of these corpora. Note that we are not singling these papers out in order to criticize; their approach of using one corpus to determine lexical frequency is standard practice in psycholinguistics.

A few papers were also reviewed to see whether the difference between corpora is discussed in the literature. Issues with corpora use are being discussed, although not necessarily in terms of frequency studies. For the most part, the only studies that discussed the frequency of corpora were those that were either creating a new corpus and thus looked at multiple corpora (Van Heuven, et al. 2014) or were specifically comparing two corpora (Burgess & Livesay 1998).[3]

We chose COCA for this replication because the frequencies were more representative of the Canadian participants used in the experiment. But, as we saw above, the choice of corpus changed the outcome of the analysis: with one corpus, we found a significant result and with another corpus we did not. We infer from this experience that the choice of corpus could have the same effect on any experiment that makes use of frequency. This brings into question the findings of every experiment that made similar assumptions to us and MBR. This finding is not necessarily surprising in the abstract, but this study provides a concrete example of these assumptions playing out.

5.2 Replication

In Psychology, the lack of replication has been well documented as a crisis (see, for example, Diener & Biswas-Diener 2016). We assume here that psycholinguistics is susceptible to the same criticism. Of course, this lack of replication calls into question the validity of the results of any one experiment and then any experiments that assume the results of a previously un-replicated experiment. The field quickly turns into a very delicate house of cards. The reason for this lack of replication is somewhat obviously due to the lack of prestige that replication carries for publication (Porte 2013) and by extension the tenure process. We offer no solutions to the crisis here. Rather we pause here for a moment to remind everyone why replication is so important. We conducted two analyses using two different corpora and found contradictory results. Either one of

[3] One study by Erker & Guy (2012) briefly discussed the methodological challenges in the study of frequency, one of which is the quantification of frequency. In their study, Erker & Guy (2012) discussed the relevance of databases such as the Brown Corpus (Francis & Kučera 1979), and CELEX (Baayen, Piepenbrock & van Rijn 1993) for specific speech communities.

these corpora could have been chosen and used alone and the findings would have been taken at face value.

It is not our purpose here to refute the findings of MBR. Our purpose is to show how very important replication is to the field. We argue that what we have done here is offer a concrete example that reminds us that replication is important enough to the reputation of our field and its claims of veracity that we find a way to prioritize replication studies.

5.3 Outcomes

Most of our discussion to this juncture has been about experiment design and the meta concerns of replication (as this is the purpose of our replication project). However, it is worthwhile for a moment to reflect on the possible world where this replication was conducted for the purpose of refuting or confirming the claims in MBR. Due to the nature of our replication, there were a few possible outcomes we could have expected and thus a few possible narratives we could have told to explain our results (Table 3). We can imagine that, if both our replication and our reproduction had returned negative results, as MBR had, we would have proclaimed MBR confirmed. If both our replication and our reproduction had returned positive results, we would have written a paper declaring MBR refuted. Table 3 shows the three logical possibilities from our study.

Table 3. Possible outcomes of experiment results.

	McCormick et al. (2009)	**COCA Replication**	**CELEX Reproduction**
Outcome A	No effect of frequency	No effect of frequency	No effect of frequency
Outcome B	No effect	Effect	Effect
Outcome C	No effect	Effect	No effect
Outcome D	No effect	No effect	Effect

Outcome A is what we would have expected, prior to beginning our study. We had no reason to think that the findings of MBR were not correct. As there was no effect of lexical frequency on RT in the original study, it seemed plausible to assume that a replication would produce the same results, whether a different corpus was used or not.

If the replication and reproduction had produced *Outcome B,* it would have meant that the difference between MBR on one side and our experiment on the other was most likely attributable to a difference in the participants or that one of the two studies (MBR and ours) had false results. If we are honest, we would have probably reported that we had refuted MBR. It would have been the lead story of our paper.

If we were to have had different results depending on which corpus was used (COCA or CELEX), *Outcome C* seemed to be most likely and would have been the easiest to interpret. The conclusion we would draw if the two CELEX experiments found no effect but the COCA experiment did is that it showed that COCA was the better corpus to represent our participant pool and thus the frequencies it contained better represented the frequencies we were testing for. Again, the lead story here would be a refutation of MBR and then a secondary story would be the celebration of COCA.

We did not find any of those easily interpretable results though. We found *Outcome D*: when we used the COCA frequencies we found no effect of frequency, but when we used the CELEX frequencies, the same ones MBR did, we *did* find an effect[4]. To be clear, if there is an effect of frequency on lexical decomposition to find, we would expect the COCA replication to find it, because MBR, using CELEX did not find the effect. We would have no reason to expect that our CELEX reproduction would not reproduce the findings of MBR, especially since the COCA replication found no effect. But, we found the opposite and so we are in a position where we are not able to interpret our results with any kind of theoretical certainty. What we have found for sure is only strong evidence that we need to be replicating.

References

Alameda, José Ramón, and Fernando Cuetos. 1995. Diccionario de frecuencias de las unidades lingüísticas del castellano. Oviedo, Servicio de Publicaciones Universidad de Oviedo.

Almeida, Jorge; Mark knobel; matthew finkbeiner; and alfonso caramazza. 2007. The locus of the frequency effect in picture naming: When recognizing is not enough. *Pschonomic Bulletin & Review, 14*(6), 1177-1182

Baayen, R. Harald; Piepenbrock, Robert; and Hedderik van Rijn. 1993. The CELEX lexical database [CD-ROM]. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Baayen, R. Harald; Ton Dijkstra; and Robert Schreuder. 1997. Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language, 37*, 94-117

Burgess, Curt; Kay Livesay. 1998. The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kučera and Francis. *Behaviour Research Methods, Instruments, & Computers*, *30*(2*),* 272-277.

Content, Alain; Philippe Mousty; and Monique Radeau. 1990. BRULEX: Une base de donnees lexicales informatisee pour le francais ecrit et parle. [BRULEX: A computerized lexical databse for the French language.] *Annee Psychologique, 90*, 551-566

Davies, Mark. (2008-) *The Corpus of Contemporary American English: 520 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/.

Biswas-Diener, Edward, and Robert Biswas-Diener. 2016. The Replication Crisis in Psychology. In R. Biswas-Diener & E. Diener (Eds), Noba textbook series: Psychology. Champaign, IL: DEF publishers. DOI:nobaproject.com.

Erker, Daniel; Gregory R. Guy. 2012. The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. *Language*, *88*(3), 526-557.

Hay, Jennifer. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics, 39.* 1041-1070.

[4] As noted earlier, the direction of the effect is also puzzling; we found a greater priming effect for high frequency primes than for low, which is not easy to interpret in light of the existing models of lexical access.

Kuperman, Victor; Raymond Bertram; and Baayen, Harald R. 2008. Morphological dynamics in compound processing. *Language and Cognitive Processes, 23*, 1089-1132.

Longtin, Catherine-Marie; Juan Segui; and Pierre A. Hallé. 2003. Morphological priming without morphological relationship. *Language and Cognitive Processes, 18,* 313-334.

Majerus, Steve; Martial Van der Linden. 2003. Long-term memory effects on verbal short-term memory: A replication study. *British Journal of Developmental Psychology, 21*, 303-310.

McCormick, Samantha F.; Marc Brysbaert; and Kathleen Rastle. 2009. Is morphological decomposition limited to low-frequency words*? The Quarterly Journal of Experimental Psychology, 62*(9), 1706-1715.

Merkx, Marjolein; Kathleen Rastle; and Matthew H. Davis. 2011. The acquisition of morphological knowledge investigated through artificial language learning. The Quarterly Journal of Experimental Psychology. i*First*.1-21

Perea, Manuel; Eva Rosa; and Consolación Gómez. 2005. The Frequency effect for pseudowords in the lexical decision task. *Perception & Psychophysics, 67(2),* 301-314.

Pierce, Jonathan W. 2007. PsychoPy – Psychophysics software in Python. *J Neurosci Methods, 162(1-2*), 8-13.

Porte, Graeme. 2013. Who Needs Replication? *CALICO Journal*, *30*(1), 10-15

Portin, Marja; Minna Lehtonen; Gabor Harrer; Erling Wande; Jussi Niemi; and Matti Laine. 2008. L1 effects on the processing of inflected nouns in L2. *Acta Psychologica, 128*, 452-465.

Rastle, Kathleen; Matthew H. Davis; and Boris New. 2004. The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, *11*, 1090–1098.

R Core Team. 2016. R: A language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Taft, Marcus. 1994. Interactive-activation as a framework for understanding morphological processing. *Language and Cognitive Processes, 9*, 271-294.

Van Heuven, Walter J. B.; Pawel Mandera; Emmanuel Keuleers. Marc Brysbaert. 2014. SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176-1190.