Predicting new words from newer words: Lexical borrowings in French

Abstract

This study addresses entrenchment into the lexicon of lexical borrowings. We search for all new lexical borrowings in a corpus of French newspaper texts and examine the frequency with which these borrowings reoccur in a second corpus of newspaper texts from about 10 years later. Lexical entrenchment emerges as depending on a variety of factors, including length in syllables, the original language of the borrowing, and also semantic and contextual factors. The dispersion of a word in the early corpus is found to be a better predictor of its frequency in the later corpus than its frequency, but both measures contribute to predicting the degree of entrenchment of a lexical item. The interaction between these two variables implies that borrowings are penalized for their burstiness.

1 Introduction

There are several ways speakers of a language can form new words. Affixation of pre-existing stems and affixes, as we see with the English suffix *-able*, is well studied. Speakers can form *hate* → *hateable* according to the same rules from which we have *love* → *love-able*; generative approaches such as those taken by (Aronoff 1976), (Selkirk 1982), (Halle & Marantz 1993), and (Ussishkin 2005) treat affixation in great detail. Blends such as *gaybie* 'gay ∗ baby', which Urban Dictionary defines as 'the child of a gay couple'[1], offer a second possibility for lexical enrichment (see (Gries 2004) for a discussion of blends in English). Further ways of creating new words, as pointed out by (Bauer 1983), include clipping, such as *convo* 'conversation', and acronym formation, such as *LDR* 'long-distance relationship'. Finally, the lexicon is enriched with borrowings like German *handy* 'mobile phone'. Because of their perceived lack of systematicity, borrowings have arguably received the least attention of all kinds of neologisms.

Borrowings can constitute a non-negligible portion of a language's lexicon. For instance, we estimate that new borrowings account for .082% of all tokens in the *Le Monde* treebank corpus of journalistic French (Abeillé *et al.* 2003) from 1989-1992. In other words, for every 1,000 words a reader of *Le Monde* encounters, it is very likely that she will run into a new borrowing.

Yet many of these borrowings could be nonce, or one-time, uses. To determine which borrowings are actually becoming entrenched into the French lexicon, we queried the online archives of *Le Figaro*, another full-coverage French newspaper, for the years 1996 -

[1] `http://www.urbandictionary.com/define.php?term=gaybie`

2006. If we have the frequencies of the borrowings at two different time periods, can we predict whether or not a lexical borrowing will "survive" and become entrenched into a RECIPIENT LANGUAGE, the language into which the borrowings enter? Theoretical morphology is well developed for neologisms formed by affixation. Yet the internal structure of a borrowing in a recipient language is largely monomorphemic, which complicates matters for theories wishing to derive new words from extant lexical entries. Since borrowings do not conform well to these analyses, the tendency is to treat them on a case-by-case basis, if at all.

Although in principle an infinite number of borrowings are possible, the borrowings that do survive are a highly constrained subset of the possible borrowings. Our working hypothesis is that borrowings do not occur indiscriminately, but are constrained by a range of different factors. Firstly, the initial DONOR LANGUAGE of the borrowings could co-determine entrenchment. For example, borrowings from a prestigious language like English could be more likely to undergo lexical entrenchment than borrowings from a less prestigious language like Polish. Furthermore, the FREQUENCY of a borrowing at a given time could be an influential predictor about the borrowing's entrenchment in the language at a later date. A borrowing's DISPERSION – the number of text chunks a word occurs in if a text is divided into several sub-parts – might also be a relevant predictor. Since shorter borrowings require fewer processing resources, we hypothesize that the LENGTH of the borrowing will be inversely related to its degree of entrenchment. We also examine a semantic factor, the SENSE PATTERN (monosemy versus polysemy) of a borrowing in the recipient language, as well as the CULTURAL CONTEXT in which the borrowing is used: whether or not the borrowing in a particular context refers to the culture of the language of the borrowing (e.g. a Russian borrowing when describing Russia, as opposed to describing China).

We are not the first to argue that a new word's frequency and dispersion are important in predicting their entrenchment. (Metcalf 2004) proposes the *FUDGE* hypothesis: new words that are FREQUENT, UNOBTRUSIVE, have a great DIVERSITY of users, that GENERATE new usages and meanings, and that refer to ENDURING concepts are more likely to become entrenched than those that are not. These are common-sense factors; in order to develop this hypothesis into a full-fledged theory, we need to test and quantify each factor against empirical data. For some of the *FUDGE* factors, this can be easily done. Metcalf's FREQUENCY and DIVERSITY correspond well to our FREQUENCY and DISPERSION measures, respectively. It is difficult to know whether borrowings are truly unobtrusive in the recipient language's lexicon: in fact, because they stem from a different language, with a different phonology and a different morphology, they may well stand out. Borrowings constitute a robust method of lexical enrichment for many languages, including French. A concept's ENDURANCE is contingent on so many cultural and societal issues that it is impossible to include in the present study.

This study extracts initial new borrowings from the *Le Monde* corpus (Abeillé *et al.* 2003) from 1989-1992 (Section 3.1) and then queries the online archives of *Le Figaro* for the same borrowings during 1996-2006 (Section 3.2), taking frequency in this second corpus as an indicator of entrenchment into the French lexicon. Based on the predictor variables of FREQUENCY, DISPERSION, LENGTH, the donor LANGUAGE of the borrowing, and the borrowing's SENSE pattern and cultural CONTEXT, we model the frequency information

of the *Le Figaro* corpus (Section 4) and find our model to explain a high proportion of the variance in the *Le Figaro* frequencies. In Section 4.2, we discuss model results, while Section 5 offers a general discussion of the findings. First, however, we go into more detail about relevant definitions and the scope of the study.

2 Lexical borrowings: definitions and classifications

This study aims to determine what factors promote entrenchment of new lexical borrowings from various donor languages into French, the recipient language in question. This use of the term BORROWING is consistent with that of Thomason & Kaufman (1988), who characterize a word as a borrowing only when fluent speakers of the recipient language adopt the lexical item from the donor language.

For the purposes of the present study, a LEXICAL BORROWING is defined as a lexical item (lemma) from a donor language satisfying the following criteria:

1. the (approximate) form and meaning are copied from donor to recipient language, without adaptation to French morphological and graphical conventions;

2. the borrowing's form-meaning correspondence is not yet found in a particular French dictionary.

The first criterion allows for the proper amount of imprecision with which to take into account speakers' probable perceptions of lexical borrowings. For example, (Thogmartin 1988) notes that his respondents offer *week-end* and *parking* as examples of English borrowings in French. Yet the form of *week-end* is not exactly the same in the donor language, since in English there is no hyphen. And the meaning of *parking* is not exactly the same in French and English: French *parking* is equivalent to English *parking lot*. A more significant divergence from meaning in the donor language concerns the German borrowing *handy* 'mobile phone': Germanophones perceive this word as an English borrowing (Erling 2004:132), but the meaning of *handy* in English is quite different from this. Since the goal of the current work is to capture speakers' knowledge about their language's lexicon, it seems appropriate to disregard these differences in form-meaning transfers and to still accept these examples as borrowings under the above definition.

Under the first criterion, the written form of the lemma must have been adopted more or less 'as is'. For example, *dérégulation* does not constitute an English borrowing according to this definition: the written form changed from English *deregulation* to *dérégulation*. Although the French form is approximately the same as the English, the French form follows French morphological and graphical conventions. That is, the affixes *dé-* 'de-' and *-tion* '-tion' are productive French affixes, and the base resembles the French *règle* 'rule'. Since French speakers most likely do not know that this word is a calque from English, we want to exclude this word from our study on borrowings.

The second criterion justifies using a term other than LOANWORD, for which typically only the first criterion applies. It also follows the initial requirements for neologisms of (Baayen & Renouf 1996). While a dictionary is not a perfect representation of speakers' lexical knowledge, often lagging behind the spoken language, verifying the presence of a word in a dictionary does represent a concrete way to measure lexical entrenchment. The

dictionary used in the present study is the *Trésor de la langue française informatisé*, or *TLFi* (Dendien & Pierrel 2003), available online at `http://atilf.atilf.fr/tlf.htm`. This is one of the more exhaustive French dictionaries. This dictionary is somewhat conservative in adding new words. Thus borrowings in the *TLFi* tend to be well entrenched.

Under this definition, transliterations into the Roman alphabet, where no apparent adaptation to French morphology occurs, also count as lexical borrowings. For example, Russian *glasnost* 'transparency' is obviously Romanized. Yet *glasnost* does not exist in a French dictionary, and it is a transliteration of a Russian term, as opposed to a translation (*transparence*) or a written Gallicization such as *glasnoste*. Lexical borrowings are not limited to words and may include idiomatic and multi-word expressions such as *last but not least* and *res nullius* 'unowned property' (Latin).

Proper nouns such as names of places or entities, titles of address, movie and book titles, and product names were excluded from the study. Also excluded from the study are any morphologically inflected or derived realizations of lexical items already existing in the *TLFi*. For example, we eliminated an attestation of *self-made men* from the *Le Monde* corpus, since it is the plural form of *self-made man*, which does exist in the *TLFi*. Finally, only exact matches were counted in the *Le Figaro* data, and near matches, such as *self government* when searching for simply *government*, were excluded.

3 Lexical borrowings, then and now

3.1 Then (*Le Monde*, 1989-1992)

The version of the *Le Monde*, or T1, corpus that we initially received from Anne Abeillé has 645,746 tokens[2]. Foreign words are labeled as such in the T1 corpus, but often, the part-of-speech tags in the corpus did not correspond to the definition of a lexical borrowing given above. The labelling of foreign words was non-systematic. For example, some proper nouns, such as *Aston-Martin*, were labeled as foreign words, and some words labeled as foreign words were in fact in the *TLFi*, such as *manu militari* 'by force' (Latin). These words are not unfamiliar to (at least some) Francophones, and it is probably best not to discuss them in terms of foreignness. To estimate the rate of lexical borrowings not labeled as foreign words in the corpus, we manually examined 431 random sentences (4% of the corpus). According to our definition, only three out of 14 lexical borrowings in this sample were labeled as foreign words in the corpus. A second pass through the corpus was necessary to find more borrowings.

This second pass to capture the borrowings not labeled as such was done in a series of three steps. The first and principal step of the procedure consisted in examining letter combinations of low frequency in 'native' French words and of relatively higher frequency in other languages. For example, we looked at all words with the letters <w> or <qi> in them. These letters or letter combinations are not common in French words, and they often tell of recent foreign origins. In addition, letters with certain diacritics, such as <ö>, do not exist in French, so any words with these written forms are certainly foreign. The derivational and inflectional morphological paradigms of other languages were also queried. For example, the superlative form of Spanish adjectives has the suffix *ísimo*;

[2]We have the 2006 version of the corpus. The corpus is updated from time to time.

Table 1: A breakdown of tokens and types according to languages.

Language	Types	Tokens
English	93	145
Spanish	12	35
German	7	60
Russian	7	18
Italian	6	10
Latin	5	5
Hebrew	4	4
Polish	1	1
Dutch	1	1
Finnish	1	1
Portuguese	1	1
Total	138	281

hence *ísimo* and *isimo* were queried. Second, the context around these borrowings was examined, on the assumption that borrowings might occur in clusters. Lastly, if step 2 led to any new borrowed words, we added the relevant written forms of these borrowed words to the list of infrequent written forms in French and began a new search for unusual letter cominations. In total, 94 letter combinations comprised of letters or letter combinations of low frequency in native French words were queried.

The combination of initial corpus labelings and this search method yielded 281 borrowed tokens and 138 borrowed string types in the T1 corpus. Borrowed types and their frequencies are given in appendix A. Table 1 presents a breakdown of tokens and types by language. It illustrates the extent to which English dwarfs all other languages for borrowings. For statistical analysis, we therefore contrasted English (93 types) with the set of other languages (45 types).

Two types, *deutschemark* 'Deutschmark'[3] and its plural *deutschemarks*, were extremely frequent, with 25 and 24 attestations, respectively. This means that one lemma comprises 17.5% of the borrowings found. Hence we perform analyses with and without this frequent lemma in the dataset to see if it is unduly biasing the results.

Our method retrieved 281 borrowings from the T1 corpus. An estimate of the number of borrowings in the corpus would allow us to evaluate the method's RECALL, i.e., the proportion of borrowings detected to the total number of borrowings in the corpus. By manually examining a subsection of the corpus for borrowings, we can estimate the recall of the method proposed as well as the number of borrowings in the corpus.

Above, we mentioned that we manually examined 431 sentences for borrowings. The method for detecting borrowings, in combination with the initial labelling of the corpus, found nine of the 14 borrowings in these sentences. Undetected tokens were *dirham*, *dirhams*, *nairas*, *majors*, and *eurobag*, where the first three of these are foreign currencies. We thus have a recall of $9/14 = 64.3\%$. We can construct a confidence interval for the method's recall, and from this we can estimate how many borrowings are in the corpus.

[3]The term in the *TLFi* for the former German currency is *mark*.

The following score confidence interval is taken from Agresti (2002:15-16):

(1) $(\hat{\pi} - \pi_0)/\sqrt{\pi_0(1-\pi_0)/n} = \pm z_{\alpha/2}$.

With $\hat{\pi} = .6429$ (9/14), $n = 14$, and with a critical value of $z_{\alpha/2} \approx 1.96$ for a 95% confidence interval, we can say that the recall of the method is between 38.76% and 83.66%. Dividing the number of borrowings obtained by both upper and lower bounds of the proportional recall gives us bounds for the number of borrowings in the corpus. Hence the 95% confidence interval for the number of borrowed tokens in the corpus is between 336 and 724. The midpoint of these two figures is 530; dividing this number by the total number of words in the corpus, 645,746, gives an estimated percentage of words in the corpus that are borrowings: 0.082%.

In order to evaluate the viability of these lexical borrowings, we consulted a second corpus sampled 10 years later.

3.2 Ten years later (*Le Figaro*, 1996-2006)

We chose the online archives of the *Le Figaro* as the T2 corpus for several reasons. First, no other large corpus of a comparable genre – general journalistic French, with a national and international perspective – was available for the time period desired. Second, as opposed to *Le Monde*, the archives of *Le Figaro* systematically provide the queried word in context. The *Le Figaro* archives date from 1996 to the present, and, wanting to have the maximum spread possible in frequencies at T2, we queried the archives from November 1, 1996 through December 31, 2006. Ideally, we would examine the borrowings continuously from 1989 - 2006 as opposed to separating the T1 and T2 corpora by ten years. This would allow us to see much more fine-grained trends in lexical entrenchment. However, no such continuous data was available for French.

We manually examined up to 200 *Le Figaro* occurrences of each borrowing to control for its sense in context. For example, one of the borrowings was *bush*, in the sense of the Australian outback, but the query results for this borrowing overwhelmingly refer to the former American presidents, so we do not want to count the latter occurrences at T2. If a borrowing had more than 200 occurrences, we estimated the total number of borrowings with the same sense by manually examining a subset of the occurrences[4], and then estimating the number of occurrences with the same sense from the inspected sample.

When we started querying the *Le Figaro* archives, the maximum number of responses returned was 1000. Ten borrowed types gave 1000 responses, which we must take as a lower bound on the total number of borrowings at T2. We also estimated the frequencies of these borrowings, albeit in a different way. This was done by manually examining up to 200 tokens of the most recent borrowings for sense, and then, assuming a uniform distribution across the 10-year period, we estimated the total number of occurrences for

[4]The exact size of the subset was determined by dividing the total number of occurrences by a divisor between two and five, in order to get the closest number of occurrences to 200 as possible. The estimated number of total borrowings was then found by multiplying the number of borrowings found in the subset by the original divisor. For example, if there were 865 occurrences of a borrowing, 173 occurrences were examined, because $865/5 = 173$. If 112 of these borrowings corresponded to the sense seen in the original corpus, then the estimated number of borrowings is $112*5$, or 560.

the entire 10 years from this sample. For example, if there were 1000 occurrences of a borrowing returned by the *Le Figaro* archives, we examined the most recent 200 of these. Let us say that, of these 200, 125 corresponded to the sense we were seeking, and let us also say that these 125 senses spanned a period of two years from 2004-2006. Since there are five two-year periods in the T2 corpus, we would then assume that this borrowing occurred with the same sense about as frequently in the other four two-year periods, which means that the estimated total number of times this borrowing with this sense occurs in the T2 corpus would be 125∗5, or 625 times. Unfortunately, over the course of querying the corpus, the maximum number of responses was limited to 300. This only affected results for three borrowings, namely, *a contrario* 'on the contrary' (Latin); *apparatchik* 'organization or party official' (Russian), and *board*. For these borrowings, the estimation technique was the same as for the borrowings for which the lower bound of 1000 responses was returned.

If the overwhelming majority of the occurrences were for a different sense, the borrowing was excluded from the T2 study. This was the case with *bush*. Unfortunately, the *Le Figaro* query interface does not differentiate between capital and lower-case letters, nor between accented and unaccented characters. Hence a query for *deregulation* also yields results for *dérégulation*. Seven borrowings were therefore discarded from the follow-up study.

The lower panel of Figure 1 shows the estimated probability density function (PDF) of the response variable, T2 frequency. The borrowings detected at T1 have a bimodal distribution at T2. This distribution suggests a difference between infrequent, non-entrenched borrowings that are part of the left bulge of the density distribution, and the well-entrenched borrowings on the right part. Examples of infrequent borrowings are *Karenztag* 'sick-day leave' (German) and *french mafia* [sic], while examples of entrenched words are *high-tech* and *nomenklatura* 'high-level government officials [under Communism]' (Russian). This result echoes the findings of (Baayen & Lieber 1997), who argue that the bimodal distribution of Dutch words with a particular prefix shows a difference between frequent, well-entrenched lexical items and infrequent nonce formations using the prefix.

In our data, a difference of ten years between corpus dates is sufficient to show patterns of entrenchment. In other words, it is unlikely that a correlation between frequencies at T1 and T2 stems from sampling from the same distribution. The upper panel of Figure 1 gives the estimated PDF of the same borrowings at T1. Comparison of the upper and lower panels indicate that the T1 and T2 distributions are quantitatively different. As we move from T1 to T2, we see a pattern of entrenchment for many borrowings: much of the probability density at T1 is concentrated at the leftmost (low-frequency) bulge, while at T2 the rightmost, high-frequency bulge is more prevalent. For example, *cash flow* has moved from the leftmost bulge at T1 to the rightmost at T2. This visual inspection is confirmed via a two-sample Kolmogorov-Smirnov test ($D = 0.523$, $p < 0.001$): it is improbable that the T1 and T2 distributions of borrowings are the same.

4 Modelling the entrenchment of lexical borrowings

We studied the T2 frequencies with the help of a multiple regression model.

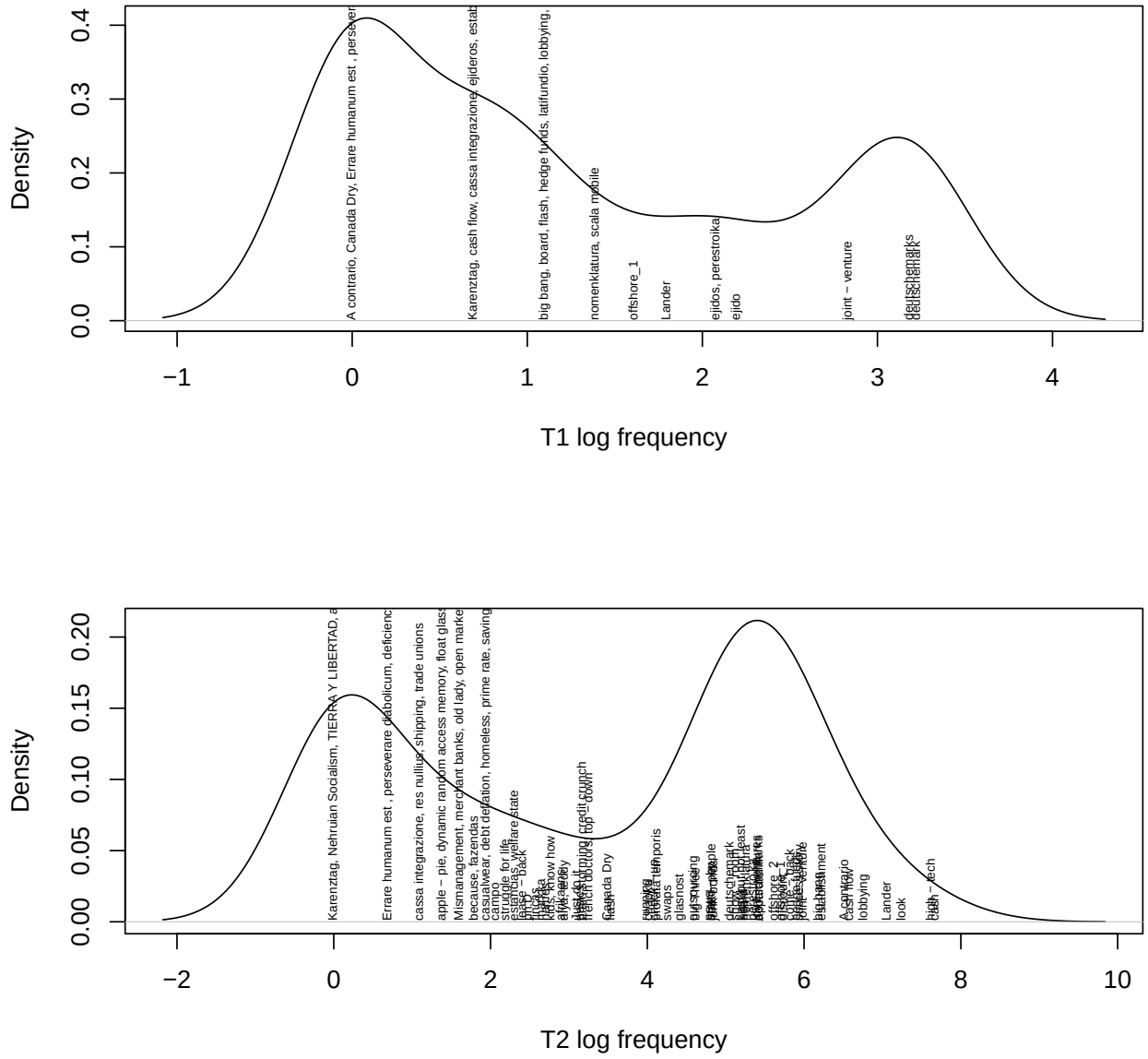

Figure 1: Estimated PDF for logarithmic frequencies of borrowings at T1 (above) and at T2 (below). Borrowed types are given at their respective log frequencies, and the T2 distribution is the response variable. Note that a density function will automatically smooth across a discrete distribution; hence frequencies smaller than 0 are given non-negative probability.

4.1 Predictor variables

Predictor variables fall into two categories: predictors representing properties that are intrinsic to a borrowing, and properties that are contextual in nature. The intrinsically defined variables examined in this study are LENGTH, SENSE PATTERN, and donor

LANGUAGE of the borrowings. The locally defined variables are FREQUENCY at T1, DISPERSION at T1, and cultural CONTEXT.

4.1.1 Length

The borrowings in the T1 corpus vary with respect to length, with one of the shorter being *names*, while the longest is *Errare humanum est, perseverare diabolicum* 'To err is human, to persist [in so doing] is diabolical' (Latin). Some researchers (e.g. (Pergnier 1989)) hypothesize that short, "punchy" borrowings are more likely to become entrenched than longer borrowings. This hypothesis makes sense from a processing perspective: shorter words could be easier to process. (New *et al.* 2006) found that for all but the very shortest words, processing times increase with length. If length has an effect on processing times, it could have an effect on lexical entrenchment. Longer borrowings might take longer to process, and over time, these words might die out due to their elevated processing costs, which are already high due to their novelty and foreignness.

There are several ways to quantify the length of a borrowing: one can count the number of letters, the number of syllables, or the number of morphemes. These different measures are highly inter-correlated: (New *et al.* 2006) mention a correlation of $r = 0.81$ for the number of letters and the number of syllables for English words. The number of morphemes is a measure that is difficult to apply for borrowings, since borrowings do not fit into the morphology of the recipient language. We hence opted to approximate length by number of syllables. In New et al.'s multiple regression model for predicting reaction times of English words, the number of syllables has a higher standardized coefficient than the number of letters. The number of syllables is given in the variable LENGTH, which was log-transformed to reduce the effects of outliers.

4.1.2 Sense pattern

Consider the borrowing *news*, which is not attested in the *TLFi*. This word can indicate a *newsletter* (a sense attested at T1 and T2) or *new information*, a sense attested only at T2. Another existing sense for *news* could help or hinder this borrowing to survive. From a utilitarian perspective, we could argue that the increased range of denotata of a polysemous word could make it be used more often. Since it can be used more often, it has greater benefit to speakers, and therefore is more likely to become entrenched.

From a processing perspective, the polysemy of a lexical item could facilitate quicker access, or it could confuse the reader, forcing him or her to disambiguate between senses. Results from the processing literature in this area are highly ambiguous. On one hand, for shallow processing such as lexical decision tasks, polysemy seems to be helping access ((Piercey & Joordens 2000) and (Rodd & Marslen-Wilson 2002); see (Rodd *et al.* 2004) for a connectionist model), but on the other, for deeper processing, polysemy might be detrimental to lexical access ((Klein & Murphy 2001); (Piercey & Joordens 2000)).

A lexical borrowing was determined to be polysemous if it has at least one other related sense elsewhere in the T1 or T2 data. It would have been ideal to examine the sense pattern at T1 only, but due to data sparsity at T1, this was not possible. Many of the borrowings only occur once in the T1 corpus, and other electronic resources from a similar genre from the same period are not available. We feel that the importance

of investigating sense pattern as a factor influencing lexical entrenchment surpasses this methodological concern.

For some new polysemous borrowings, an existing sense was already present in the *TLFi*. For the purpose of this paper, a SENSE is defined as a first-level sub-entry of a lexical entry in the *TLFi* dictionary[5]. This dictionary-based definition is fairly intuitive: homonyms are considered different lemmata, but related senses are noted as different sub-entries of the same lemma. For example, a *pack* has one entry with two immediate sub-entries: it is both a 'a number of individual units packaged as a unit' or 'a group of eight forwards in a rugby game', and not simply 'a group of teammates', which is the approximate definition it has in the T1 corpus. The variable SENSE is binary, and its values are MONO (monosemous) and POLY (polysemous).

4.1.3 Donor language

Another factor that could play a role in the entrenchment of a borrowing is the donor language from which it stems. Much literature discusses the privileged status of English borrowings among borrowings in French (see (Etiemble 1964) and (Hagège 2006) for examples of highly popular general-audience works on this topic). Yet for all of the claims about the reasons for and effects of English borrowings in French, most research ((Picone 1996), (Pergnier 1989), (Rey-Debove 1987), etc.) examines only English borrowings and not lexical borrowings from other languages. Although the number of English borrowings detected at T1 is far greater than the number of borrowings from any other language, it is not obvious whether these borrowings will also occur more at T2. In other words, we would like to know whether English borrowings have the same probability of entrenchment as borrowings from other languages. Donor language is investigated in the variable LANGUAGE. As discussed in Section 3.1, this variable is binary (ENG, i.e. English, vs. NON-ENG, non-English) due to the distribution of the borrowings in Table 1.

4.1.4 Frequency and dispersion

The effect of frequency is extremely prominent in language and has been studied most recently with respect to processing cost (see (Bybee & Hopper 2001) for a general overview). In the context of our study, we ask whether a borrowing's frequency at T1 will predict the borrowing's frequency at T2. In this study, a borrowing's frequency is its overall log count in the T1 corpus, since word frequency is perceived on a logarithmic scale (see, for example, (Howes & Solomon 1951) and (Oldfield & Wingfield 1965)).

Frequency information is given in terms of counts in a corpus of texts, yet dispersion can also measure the degree of entrenchment of a lexical item into the lexicon. Which of these two measures is a better indicator of entrenchment? In other words, if we see a word with a frequency of 2 occurring in two different text chunks, and another word with a frequency of 6 which only occurs in one text chunk, can we say which word,

[5]A stipulation of this definition is that it includes items that are only in the dictionary as fixed expressions but that are used in a free context. The only example of this type of polysemy in our findings concerns the word *cash*, where the *TLFi* lists this item only in the expression *payer cash*, yet it exists freely in our findings ('...30 % du prix devant être versé en *cash*', here with our italics. The data format of the corpus does not specify whether the borrowings were originally italicized.)

if either, is more likely to be entrenched into the lexicon? Furthermore, perhaps it is possible that the presence of one of these measurements obviates the need for the other. We have no a priori opinion about which measurement, if either, will be more effective than the other. However, we suspect that these two measures will complement each other, and that knowledge of both will only improve model accuracy in predicting lexical entrenchment. This hypothesis is supported by the work of (Gries 2008), who proposes that both frequency and dispersion measures should be given when discussing cognitive entrenchment of lexical items.

Ideally, we would measure dispersion using articles in the T1 corpus as different text chunks. Since an article is a cohesive text unit, this division reflects speakers' exposure to the borrowings well. However, this was not easily possible given the format of the T1 corpus.

The T1 corpus is divided into 44 sub-corpora. Since the literature on the corpus (Abeillé *et al.* 2003) does not specify to what a sub-corpus corresponds, we determined upon inspection that several articles comprise a sub-corpus, and that almost all of the time one article was contained in one sub-corpus. In the absence of clear boundaries between articles, then, a sub-corpus is an ideal unit for obtaining dispersion measurements. In using the pre-defined sub-corpora, we hope to approximate the essential of what a dispersion measurement aims to capture, i.e. how many times the borrowings are used in different articles, while minimizing the amount of manual inspection that would be necessary to divide the corpus into articles. The 44 sub-corpora range in size from 6288 to 18594 words.

Unsurprisingly, frequency and log dispersion are correlated in our data ($r_s = 0.926$). Predictor variables with such high correlations can lead to a spurious model, so we regressed log frequency on log dispersion and took the residuals of this model as the FREQUENCY variable. This variable gives all frequency information that is not already taken into account by the DISPERSION variable, and so these two variables are independent. The correlation between original frequency and our variable, residual frequency, is $r = 0.272$, which is significant at $p < 0.001$. That is, a substantial portion of frequency information is already taken into account with the DISPERSION variable, but what is not is positively correlated with T2 frequency.

4.1.5 Cultural context

The context in which a borrowing occurs could provide a clue about its degree of entrenchment. A borrowing can occur in a context referring to a culture that is typically associated with use of the language from which the borrowing stems, or it can occur outside of that cultural context. We call the former RESTRICTED cultural contexts and the latter UNRESTRICTED cultural contexts. An example of a restricted cultural context is when using the Russian borrowing *perestroïka* 'economic restructuring' when referring to Russia or the former Soviet Union, while using it to refer to France represents an unrestricted cultural context. Using a borrowing in an unrestricted context could imply that the borrowing is less anchored to a particular culture, and that the borrowing could take on a more general meaning. This relative lack of restrictions of the use of the borrowing could correlate with lexical entrenchment.

Continuing with our example, in French, *perestroïka* occurs 210 times in the T2 cor-

pus, in contexts referring not only to the former Soviet Union, but also to Iran, Syria, and even France and the European Union. There are four attestations in which this borrowing refers to an economic restructuring of a democratic nation; one occurrence even refers to a *perestroïka médicale*. On the other hand, in the *New York Times* for the same date range, *perestroika* occurs 174 times; this number is arrived at using the same techniques of estimation used in Section 3.2. In the *New York Times* data, all but one of the occurrences refer to economic restructuring of a non-democratic regime. Perhaps in French then the meaning of this term is something closer to 'economic restructuring' than the term's meaning in English, where it still seems strongly anchored to the former Soviet Union, former Communist countries, and, to a lesser extent, other non-democratic regimes. The widened scope of reference of the French term, as opposed to the English, and its corresponding greater frequency at T2 is perhaps no accident.

Flaitz (1988:87) touches upon the notion of cultural context in her study on English borrowings in the French press when she notes that

> The words *success story*, for example, appear. . . in an article about the life and career of Lee Iococca, the American head of Chrysler Corporation. Had the article been focussed on the life and career of François Mitterrand, the anglophone phrase *success story* would have incited much well-deserved criticism.

Here *success story* is used in a restricted context when discussing the American Lee Iococca, but if it were used to describe François Mitterrand, it would be an unrestricted context. The above passage implies that the range of possible cultural contexts is a salient feature of a lexical borrowing. The current study translates this intuitive concept into a criterion for measuring the degree of a borrowing's lexical entrenchment in the recipient language: it is perhaps the entrenchment of the borrowing into the language, observable in the unrestricted use of a borrowing, that would incite the criticism. We predict that a lexical borrowing in an unrestricted context in the T1 corpus, such as *success story* when discussing François Mitterrand, will reflect a higher degree of lexical entrenchment in French and hence a higher frequency at T2.

Determining the cultural context of a borrowing was done manually, as no method for automatically determining cultural context was available. Use of a borrowing in direct connection with a culture in which that language is spoken was considered a restricted cultural context, while use of it outside of that context was considered an unrestricted context. The majority of cultural contexts were classified fairly easily. For example, contexts discussing British or American companies' *junk bonds* would be restricted, while contexts discussing *junk bonds* from French or Italian companies would be unrestricted.

However, some classification difficulties arose when discussing international contexts. If the context concerned multi-national bodies such as the European Union, we classified these contexts as unrestricted. At present, it is an open question whether these contexts correspond to the definition of a culture typically associated with a particular language. When seeing borrowings in the context of international agreements between countries that could trigger a restricted cultural context, we classified the borrowings as restricted. For example, a *joint-venture* between British and Japanese companies would have a restricted cultural context, since Britain is associated with English. Any mention of a country typically corresponding to a particular language could elicit more borrowings, and so our

guiding principle was to be conservative when classifying borrowings as unrestricted. The binary CONTEXT variable has two values, RESTRICTED and UNRESTRICTED.

4.2 Results

Ten of the 281 tokens were excluded from the study as it was impossible to gauge their frequencies in *Le Figaro*, or because their cultural contexts or sense patterns were unclear in the T1 corpus. On the remaining dataset (n = 271), we performed backward variable selection starting with main effects for all predictors and any two-way interactions, using the Akaike Information Criterion (Akaike 1974) to eliminate superfluous predictors. This yielded a model with 11 factors. We explored using a mixed-effects model (e.g. (Baayen *et al.* 2008)) on the data, but due to large differences in number of occurrences of the borrowings at T1, such a model could not be fit to the data.

The multiple regression model for predicting lexical borrowings, with the predictor variables given in Section 4.1 and log frequency at T2 as the response variable, is given in Table 2. Intrinsic properties of the borrowings are listed first, followed by locally determined properties and then two-way interaction terms.

Table 2: Linear model for predicting lexical entrenchment of French borrowings.

	$\hat{\beta}$	S.E.	t value	Pr(>|t|)
(Intercept)	3.0098	0.7177	4.19	< 0.0001
LENGTH	−0.5868	0.3511	−1.67	0.0959
SENSE (POLY)	2.1125	0.5080	4.16	< 0.0001
LANGUAGE (ENG)	−0.6973	0.5247	−1.33	0.1850
FREQUENCY	−2.8443	0.8240	−3.45	0.0007
CONTEXT (RESTRICTED)	0.5845	0.8239	0.71	0.4787
DISPERSION	1.7503	0.0902	19.40	< 0.0001
FREQUENCY∗DISPERSION	−1.9479	0.4607	−4.23	< 0.0001
FREQUENCY∗CONTEXT (RESTRICTED)	2.3039	0.8553	2.69	0.0075
LENGTH∗CONTEXT (RESTRICTED)	−1.7410	0.4635	−3.76	0.0002
SENSE (POLY)∗CONTEXT (RESTRICTED)	−1.9864	0.7363	−2.70	0.0074
LANGUAGE (ENG)∗CONTEXT (RESTRICTED)	1.8067	0.5792	3.12	0.0020

The lower panel of Figure 1 above shows that even after a logarithmic transformation, the response variable is not normally distributed. This non-normality was further indicated in an S-shaped pattern in the model's residuals, which indicates that the model finds it difficult to predict the borrowings with very small or very high frequencies at T2. This same pattern emerged after we excluded outliers. After eliminating datapoints marked as potentially overly influential from the original model and refitting the model, the same S-shaped pattern was apparent. Fortunately, a plot of the observed vs. predicted values in the original model shows general homoscedasticity of the errors. At this point, we were fairly confident that the model in Table 2 was robust. Still, to ensure that violations of the assumptions of the linear model were not leading to spurious results, we fit a non-parametric model, a random forest, to the borrowings data. We used an

implementation of a conditional permutation scheme for each predictor variable so as to avoid undue influence of correlated variables (Strobl *et al.* 2001). Information about the importance of each predictor variable in the random forest model is given in Table 3.

Table 3: Predictor variable importance for a random forest model of the borrowings data. Variable importance is given as a function of mean decrease in accuracy if the variable is held out of the model.

Predictor variable	Mean decrease in accuracy
SENSE	0.001
CONTEXT	0.033
FREQUENCY	0.492
LANGUAGE	0.505
LENGTH	1.344
DISPERSION	7.394

According to the random forest, DISPERSION is the most important predictor variable; it also explains the most variance in the linear model. The SENSE predictor variable is the least important, but is still helpful in predicting the response variable. The correlation between predicted and observed values for the random forest model is $r = 0.828$, while for the linear model we have $r = 0.832$ for an R^2 value of .6797. Since the parametric and non-parametric models are similar, we conclude that the violations of the linear model are not sufficient to warrant concern – they do not prevent the model in Table 2 from being robust. Hence all subsequent results are in reference to the linear model.

Two types, *deutschemark* 'Deutschmark'[6] and its plural *deutschemarks*, were extremely frequent, with 25 and 24 attestations, respectively. This single lemma comprises 17.5% of the borrowings found, and it is necessary to determine whether it is unduly influencing the model we propose. To answer this question, we excluded all occurrences of *deutschemark* and *deutschemarks* from the data and reran a model on the data without these types. The resulting model was very similar to the original model, and we conclude that this lemma alone is not significantly affecting the results.

Since (New *et al.* 2006) show a U-shaped curve for the effect of length on lexical decision times, we also added a quadratic term for syllable length to the model. This term was not significant: with respect to new borrowings, a linear effect of length, as opposed to a polynomial effect, is sufficient for predicting lexical entrenchment. The lack of significance of the quadratic term in our model perhaps stems from the great majority of the borrowings in the corpus falling into the medium- and long-length categories of (New *et al.* 2006). Also, we have measured length in terms of syllables, not letters. Therefore, a potential quadratic trend of length in the population is not likely to be significant in the model given the present dataset.

The final model given in Table 2 has an adjusted R^2 value of 0.6797 with a residual standard error of 1.402. To determine whether the model was overfitting the data, we performed bootstrap validation and compared the original goodness of fit statistics of the

[6]The term in the *TLFi* for the former German currency is *mark*.

bootstrap samples to the full dataset. The R^2 decreases from 0.6851 on the training set to 0.6650 on the test set for an optimism of 0.0286. This optimism is no doubt due to the small dataset, but it is not large enough to call the model into question. Based on this information, we are reasonably confident that the model is predictive: it is capturing elements of lexical entrenchment of borrowings in French.

In this model, the significant main effects are SENSE, FREQUENCY, and DISPERSION, with the number of syllables being only marginally significant at $p < 0.10$. These main effects characterize the unrestricted contexts; all significant main effects except DISPERSION enter into significant interactions with CONTEXT. The DISPERSION predictor has the heaviest weighting and highest significance. All interaction terms are significant at the $p < 0.01$ level, with the interactions between DISPERSION and FREQUENCY as well as LENGTH and CONTEXT having the greatest influence in the model. As all significant main effects are modified by interactions, they will be discussed in conjunction with their interactions.

Figures 2 and 3 plot the partial effects for predictors retained in the model; the next subsections detail these results.

4.2.1 Length

As is seen in the upper left panel of Figure 2, length enters into a significant interaction with cultural context. When the cultural context is restricted, an increase in length of a borrowing has an even more negative effect on the likelihood that the borrowing will become entrenched.

In examining the data, an explanation emerges as to why the effect of length is more pronounced when the cultural context is restricted. Of the 13 borrowings with the highest number of syllables, nine have restricted cultural contexts. Many of these borrowings refer to a specific policy in a foreign country, such as *an elastic currency*, *cassa integrazione* 'partial unemployment' (Italian), *classless society*, referring to the specific goal of John Major, the former Prime Minister of Great Britain, and *tierra y libertad* 'land and freedom' (the Spanish cry of Mexican revolutionaries). We are probably seeing this interaction because specific policies tend to have longer descriptors than more general concepts. In general, we can conclude that length probably does not bode well for a borrowing, and when length is used to denote a specific policy of a foreign country, the borrowing is even less likely to become entrenched.

4.2.2 Sense pattern

The next intrinsic property of a borrowing we examine is its sense pattern. The upper right panel of Figure 2 provides evidence for the hypothesis that in culturally unrestricted contexts, polysemous borrowings – i.e., new senses of borrowings with extant senses elsewhere in the language – are more likely to become entrenched than monosemous borrowings. While a polysemous borrowing has an extended range of denotata, borrowings in restricted cultural contexts have more limited ranges. The limited range of restricted cultural contexts could mitigate the positive effect polysemy has on a borrowing's probability of entrenchment. Monosemous borrowings, regardless of the context in which they occur, are more likely to be doomed to the margins of the lexicon.

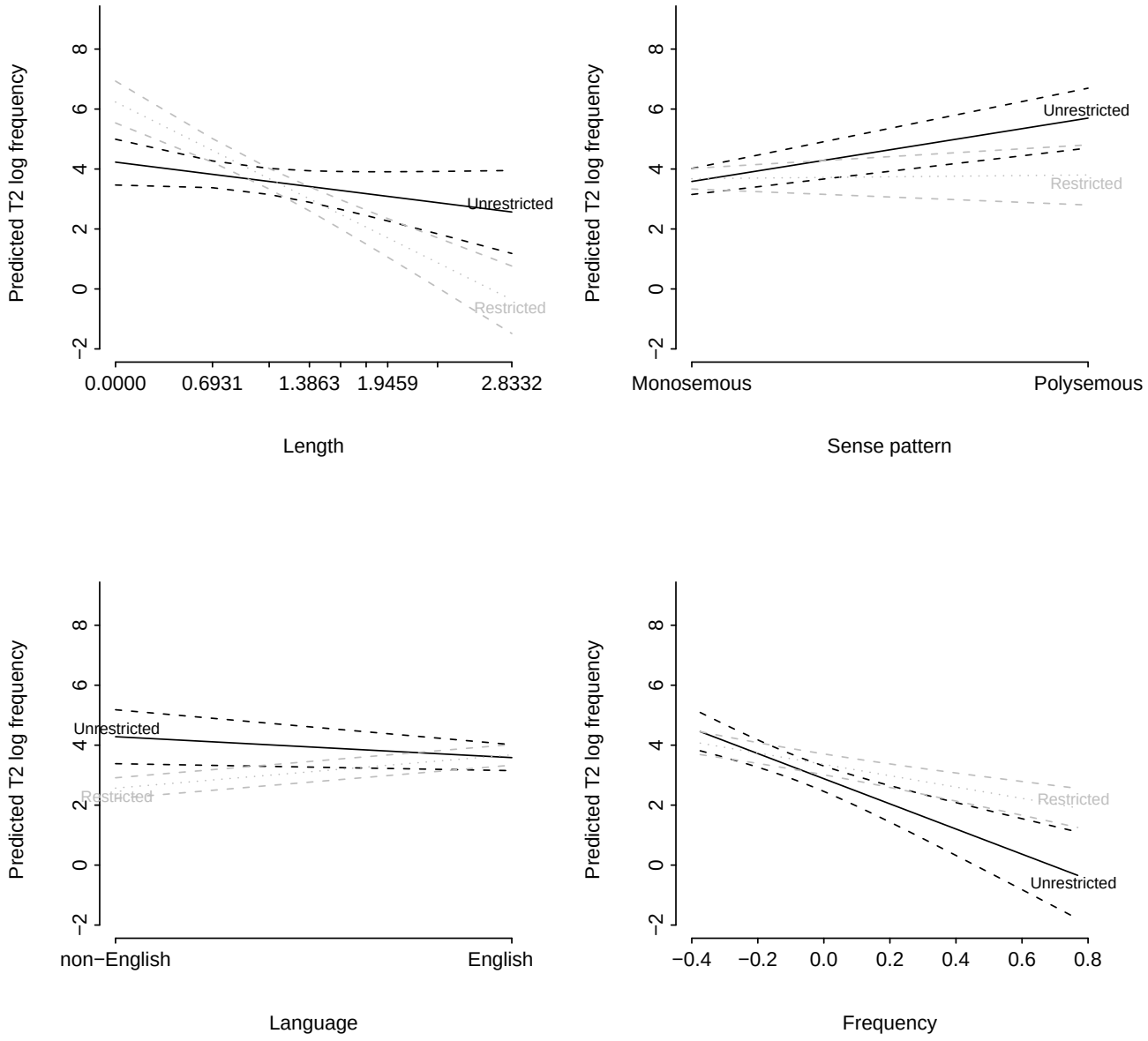

Figure 2: Effects for word length, sense pattern, donor language, and frequency by cultural context (restricted or unrestricted). Dashed lines indicate the 95% confidence intervals for each effect.

Because of the interaction between polysemy and cultural context, we argue that the utilitarian explanation is better suited to explain our findings on polysemy than the processing account. Polysemy is helpful in explaining the occurrences of the borrowings at T2, but only in culturally unrestricted contexts. Restricted contexts most likely narrow down the sense of a borrowing to such an extent that the advantage of polysemy can no longer emerge.

Recall that our definition of polysemy requires us to include the English borrowing

cash. Over the course of our research, the dictionary entry of the *TLFi* was changed from having a main entry of *payer cash* 'pay with cash', as is noted in footnote 1, to simply *cash*. Although *cash* is still only given as a part of the multi-word expression *payer cash*, it could be seen as existing in our dictionary, and hence we might want to exclude it from our study. Since it is highly frequent at T2, including this word may unduly influence the influence of polysemy on lexical entrenchment. Upon excluding *cash* from the dataset, the effect of polysemy on lexical entrenchment is slightly diminished, but still significant ($\beta = 1.948$, $p < 0.001$).

The current findings contribute to general research on polysemy. With forms of language creation other than borrowings, we see that polysemy arises spontaneously ((Steels *et al.* 2002) look at simulations of language creation with artificial agents and finds polysemy of newly formed lexical items). Polysemy is highly prevalent in natural language: according to Fuchs (1996:29), 40% of French lexical items are polysemous, while (Rodd *et al.* 2004) observe that 84% or more of relatively frequent English words in the Wordsmyth dictionary (Parks *et al.* 1998) are polysemous. The English figure is higher the French figure since more frequent words tend to be more polysemous, but we can still see that polysemy is rampant in these two languages. Because polysemy is common and is arises spontaneously, it might be advantageous in some way.

Our hypothesis is that it is easier for speakers to make a transfer of meaning than to invent an entirely new word for a given concept. For example, on the French website Dictionnaire de la Zone (`http://www.dictionnairedelazone.fr/`), a website of neologisms in French similar to Urban Dictionary insofar as users propose content and give example usages, the word *fils* is proposed to mean 'friend, colleague, pal [masc.]'. This meaning no doubt is related to the standard definition of *fils*, i.e. 'son'. It seems less difficult to create this word with this meaning than to create an entirely new word with the same meaning.

Under this hypothesis, polysemy functions as a mnemonic device to a speech community when the neologism enters the language. On popular websites proposing or documenting neologisms in English such as WordSpy (`http://www.wordspy.com/`), Urban Dictionary (`http://www.urbandictionary.com/`), and Wiktionary (`http://en.wiktionary.org/wiki/Wiktionary:List_of_protologisms`), there are very few proposals for new words that do not have at least one element in common with an extant word; this element can be morphological, phonological, and/or semantic in nature (e.g., metonymy). This shared element most likely contributes to the retention of the new lexical item. Just as sharing a form with an existing lexical item makes learning a new word easier, so does sharing semantic content. Perhaps for these reasons borrowings relying on polysemy with extant lexical items are more likely to become entrenched than other borrowings. In this way, borrowings could also provide evidence toward a shared element with an extant lexical entry giving a neologism an "evolutionary advantage" in the lexicon.

4.2.3 Donor language

Donor language did not yield a significant main effect, but did have a significant interaction with cultural context. As is seen in the lower left panel of Figure 2, for borrowings from English, the cultural context is irrelevant. However, for borrowings from other

languages, context matters: non-English borrowings occurring in unrestricted contexts are more likely to become entrenched than restricted non-English borrowings. Hence we have Russian *nomenklatura* 'high-level government officials [under Communism]', which always occurs in culturally unrestricted contexts at T1, and which occurs 190 times at T2. In contrast, a restricted non-English borrowing is *scala mobile* 'method of adjusting salaries according to prices' (Italian), with a frequency and a dispersion of 4 in the T1 corpus but only 3 occurrences in the T2 corpus.

It is possible that markedness plays a role in this distinction. Context is irrelevant for English because it is the unmarked donor language. Francophones are so used to hearing English borrowings that the cultural context in which they occur does not matter for lexical entrenchment. For marked donor languages, borrowings in unrestricted cultural contexts are more likely to become entrenched than borrowings in restricted contexts because in unrestricted contexts, there are fewer limitations on what the borrowings can refer to. That is, the results for marked donor languages support our original hypothesis about cultural context.

4.2.4 Frequency

The lower right panel of Figure 2 indicates that, regardless of context, a higher frequency is a bad omen for a borrowing. This at first sight counter-intuitive finding can only be understood when dispersion is also taken into consideration (see Section 4.2.5). The negative effect of frequency is less pronounced when borrowings occur in restricted contexts. For example, the borrowing *board*, with the sense of *board of directors*, occurs three times at T1, always in culturally restricted contexts. It has a dispersion of 1, yet it is still frequent at T2 with 184 occurrences. On the other hand, the borrowing *flash*, with the sense of *flash memory*, also occurs three times at T1 with a dispersion of 1. Two of these three occurrences are in culturally unrestricted contexts, and it only has 33 occurrences at T2.

We posit that frequency is less of a bad omen for borrowings in restricted cultural contexts because there are more cues to aid in integration into memory of these words. For example, a word such as *board*, occurring only in culturally restricted contexts at T1, is used to describe a particular company, and having a particular company tied to our memory of *board* will help us remember this lexical item. The more frequent this lexical item is, the more associations a speaker will have with it. Borrowings occurring in culturally unrestricted contexts have fewer salient associations with them, and they are hence less likely to benefit from increased frequency.

4.2.5 Dispersion

The DISPERSION main effect and the DISPERSION∗FREQUENCY interaction are solid predictors with the largest effect sizes in the model. Since the FREQUENCY variable is also significant, we can say that knowing only about a borrowing's dispersion and frequency in the T1 corpus will enable us to make a decent guess about its degree of lexical entrenchment at a later date. This finding is somewhat surprising: first, new words can be trendy, and hence frequent at a particular time, and all but forgotten some years later. For example, this can be the case with words to describe new technologies that are not adopted by

society. A query of the *New York Times* archives for *laser disc*[7], a now-obsolete precursor to the DVD, shows that 18 of the 58 results are from 2000 or 2001, which is right around the time the production of laser discs stopped. This word is fairly frequent during these years, but is only seen three times from January 2007 to August 2008.

Second, the relationship between dispersion and lexical entrenchment is not straightforward due to the interaction between dispersion and frequency at T1. The main effect of DISPERSION shows a positive correlation with T2 frequencies, but the DISPERSION∗FREQUENCY interaction has a negative coefficient value. As dispersion and frequency increase, the number of occurrences at T2 decreases.

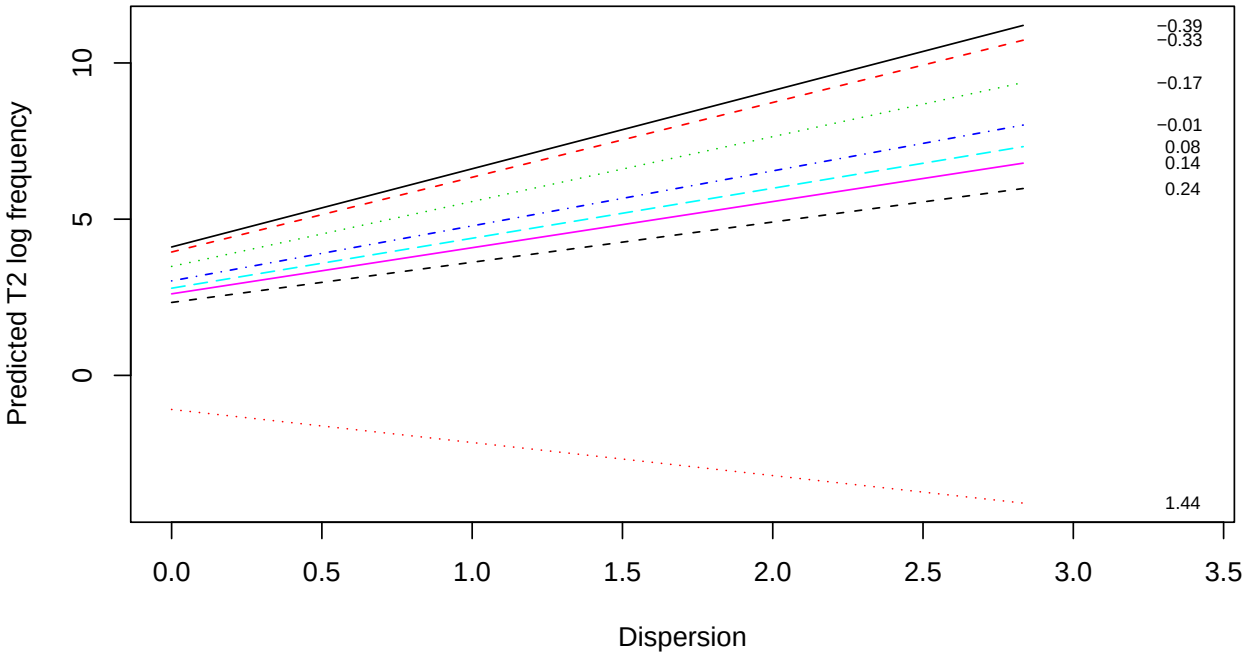

Figure 3: Interaction plot of log dispersion and residualized log frequency (in deciles given on right, with deciles 2 and 3 having equal values) on predicted T2 log frequency.

Figure 3 shows the dispersion-frequency interaction in quantiles of tenths. In this figure, residualized frequency deciles are in ascending order from top to bottom on the right side. For example, one tenth of borrowings have a residualized frequency of -0.39 or less, and 100% of borrowings have a residualized frequency of less than 1.44.

Like in Figure 2, in Figure 3 we see the same decrease in predicted T2 frequency as T1 frequency increases[8]. In this figure, the borrowings in the uppermost decile for FREQUENCY, the lowest dotted line, have the largest effect on this interaction. Borrowings in this decile have a residualized log frequency greater than 0.238. These borrowings

[7]See `http://tinyurl.com/6dnxs8`, accessed on 27 August 2008.

[8]The largest decile, shown in Figure 3 with a residual frequency of 1.44, is not given in Figure 2. This is a design principle of the `Design` package in `R` (Harrell 2001).

are frequent, yet poorly dispersed in the T1 corpus. In fact, in Figure 3, they do not extend between the 1.0 hash mark on the dispersion axis. In general, these borrowings are infrequent in the T2 corpus. Examples of these borrowings include *ejido* 'communal agricultural land', Spanish, with 9 occurrences at T1 and 0 occurrences at T2, *ejidos* (*ejido*-pl), with 8 occurrences at T1 and 1 occurrence at T2, and *flash*, with 3 occurrences at T1 and 33 occurrences at T2.

This interaction suggests that borrowings with a low dispersion but high frequencies are getting penalized for their burstiness. A word that occurs, say, nine times in one article but nowhere else in the corpus is probably more indicative of that specific article than of language use in general. Unless there is another article about the socio-political aspects of agricultural practices in Latin America, for example, the borrowing *ejido* will probably not be used in French. In contrast, a word with a frequency of 9 and a dispersion of 3 is likely to be a frequent word at T2. The actual words are article-specific, but the penalty they incur is generalizable.

In the T1 corpus, many of the borrowings in the top decile for residualized frequency do come from one large article on the socio-political aspects of agricultural practices in Latin America. Because of this, a reviewer wondered whether a burstiness penalty is more appropriate for borrowings from this article or for borrowings in general. Upon exclusion of borrowings from this article in the multiple regression model, a positive DISPERSION main effect and a negative FREQUENCY∗DISPERSION interaction term were still significant. These results suggest that all frequent yet underdispersed borrowings in the dataset, and not just those in this specific article, incur a burstiness penalty.

4.2.6 The causal nature of frequency and dispersion

From a strictly synchronic perspective, our predictor variables of FREQUENCY and DISPERSION are only diagnostics, and not causal factors. It is essential to take diachronic corpora into account when examining lexical entrenchment, because otherwise our predictor variables of FREQUENCY and DISPERSION would only be diagnostics, and not causal factors, in determining lexical entrenchment. In fact, the *FUDGE* hypothesis has been criticized as at least partially circular (Pinker 2007:308), because frequency and diversity of users are, in a synchronic approach, what one would like to explain. But these factors can be causal predictors in a diachronic approach to lexical entrenchment: the more frequent and well-dispersed a new word is, the more speakers will hear and eventually use it.

We can predict not only the frequency with which a word occurs at T2, but also whether it is still in use at T2. Crucially, a logistic regression model predicting whether the word will still be in use 10 years later is indeed significantly better than chance at predicting whether words fall out of use ($\chi^2(1) = 6.487$, $p = 0.011$). The model correctly predicts that borrowings are no longer in use in 39 cases and that they are still in use for 194 cases. Our model can effectively predict lexical change across two time periods on the basis of FREQUENCY and DISPERSION, inter alia[9].

[9]The logistic regression model has main effects for FREQUENCY ($\beta = -6.394$) and DISPERSION ($\beta = 5.033$) and interaction terms LENGTH∗CONTEXT ($\beta = -2.732$) and LANGUAGE∗CONTEXT ($\beta = 3.188$), all significant at $p < 0.05$.

4.2.7 Cultural context

Of all the predictor variables, CONTEXT has the most complex relationship with the response variable. The main effect of cultural context is not significant, yet it interacts with four other predictor variables, LENGTH, SENSE, LANGUAGE, and FREQUENCY, in ways that are perhaps not intuitive.

A rule of thumb of linear regression is that there should be 15 datapoints for every factor (Harrell 2001:61). Our model has approximately 23.4, so we are not overfitting the data. Furthermore, in all of the models explored above in Section 4.2, i.e. the model without outliers, the model without influential datapoints, and the model excluding the frequent lemma *deutschemark*, the cultural context effects remained very similar to those of the original model. It is therefore unlikely that the cultural context interactions in the model are caused by the exceptional properties of our dataset.

Our current hypothesis is that borrowings tend to first enter the language in restricted contexts and then expand their contexts to unrestricted borrowings. It follows straightforwardly that borrowings seen at T1 in unrestricted contexts are more likely to become entrenched, since they are further along in the process of entrenchment than borrowings in restricted contexts. Future work can directly test this hypothesis by examining the contexts of the borrowings at T2 of borrowings occurring in restricted contexts at T1. If at T2 we see a positive correlation between frequency and unrestricted uses of these borrowings, we can conclude that some borrowings do initially have restricted cultural contexts before the set of contexts in which they can be used is expanded.

The cultural context can be conceived of as a coarse-grained content indicator. Is content otherwise relevant to a borrowing's adoption? Does topic matter? For example, one could hypothesize that technological borrowings are more likely to become entrenched than borrowings pertaining to fashion. Unfortunately, the lack of information about the breakdown of topics in the T1 corpus does not allow for us to answer this question in a meaningful way. We are not given a distribution of articles by topic and/or by genre for the T1 corpus. This means that if we see more financial borrowings, for example, we do not know if this is because a high proportion of the corpus relates to financial articles, or if borrowings relating to financial topics really are more likely to become entrenched than borrowings pertaining to other topics. This could be a feature of upcoming versions of the T1 corpus (Anne Abeillé, p.c.), so this may well be a promising avenue for future research. Another improvement to the T1 corpus concerns the division of the corpus into articles as opposed to sub-corpora. Divisions by articles will be a feature of future versions of the corpus (Abeillé, p.c.), and working with articles could allow for enhanced dispersion counts.

5 General discussion

Given that a new word occurs in a language, what makes it likely to "survive" and become part of the lexical stock of the language? To answer this question, we examine one method of lexical enrichment, lexical borrowings, in French, a language in which new lexical borrowings are not only common but also relatively easy to single out in corpora. We look at two journalistic corpora in French from two different time periods to see if the

new borrowings found at the first time period (T1; 1989-1992) are still in use at the second time period (T2; 1996-2006). Using frequency at T2 as a measure of lexical entrenchment, we find that several factors, such as a borrowing's dispersion and frequency, whether or not a borrowing is polysemous, the length and cultural context of a borrowing, and donor language of the borrowing help to determine a borrowing's degree of lexical entrenchment into French.

Of the factors examined, we would expect that dispersion, frequency, sense pattern, and length will also be relevant to a similar study on different types of neologisms, while donor language and cultural context are specific to borrowings. Given our results, all of these factors could be relevant when examining borrowings in other recipient languages, but we do not necessarily expect the factors to act the same way as they do in French. Non-English borrowings in culturally unrestricted contexts are more likely to be incorporated into the lexicon than culturally restricted non-English borrowings. Longer borrowings are generally less likely to become entrenched. Polysemous borrowings, i.e. borrowings with extant senses in the lexicon, are more likely to become entrenched into the lexicon than borrowings with no existing senses in the lexicon. Some factors, such as dispersion, are relatively straightforward to interpret. The interactions involving context are more difficult to understand, and the explanations we offer about them remain tentative.

It remains an open question as to whether the presence of pre-existing words or phrases in the recipient language at T1 describing the same concept as the borrowing can affect a borrowing's entrenchment. If extant native words are available for speakers, this might be reflected in a lack of lexical entrenchment for the borrowing. We opted to steer clear of this potential predictor, however, because it is impractical for at least three reasons. First, it sometimes is difficult to know at T1 if the equivalent word exists in the recipient language. For example, the English borrowing *short* is given in a stock context to refer to short sellers, and the French translation of *short seller* is *vendeur à découvert*, which is given in the T1 corpus directly preceding *short*. But who knows if *vendeur à découvert* existed before this usage? Finding this out would require detailed sleuthing for each individual borrowing. This would be further complicated by the lack of electronic resources for pre-1989 dates. Furthermore, just how many times does a word have to be mentioned in the language in order for us to say it exists in the language? If we say just once, it could be quite difficult to say with certainty that an equivalent native term has never existed before in the language. Second, in our informal discussions on the meanings of borrowings with native speakers, it seems some speakers may detect a nuance of sense in the borrowing that others do not, so that the former will say there is no exact translation of a borrowing, while the latter will gladly given a native equivalent for the borrowing. Such subjectivity could complicate this avenue of research. Finally, perhaps a native phrase or multi-word expression exists to convey a similar idea, but it is considerably longer than the borrowing. All of these considerations combine to make the effect of extant native lexical items a challenging factor to examine.

For corpus and computational linguists, the principle finding of this research is likely to be the role of dispersion and frequency in modeling lexical entrenchment. These variables are both significant predictors, and the interaction between them is also predictive, with a negative effect on the response variable. This interaction penalizes borrowings in our dataset that are frequent but too bursty. Preliminary psycholinguistics findings

(Gries 2008) support the idea that dispersion is a better predictor of reaction times than frequency. This is congruent with the present result and therefore could suggest that a burstiness effect could also be present in reaction time studies. Further studies and computational applications requiring word frequency distributions would do well to examine both frequency and dispersion for the construction of predictive language models.

For linguistics, the present findings indicate that lexical entrenchment of borrowings is predictable to a surprising extent. This is fortunate, since in rule-governed morphology, generative works give insight as to whether a form is possible or not. Productivity measures à la (Baayen 1992) assign a probability to an affix or a lexical process of occurring in new lexical items. Neither of these approaches provide insight as to whether the form will become entrenched in the language. We offer the present study as a first step in bridging the gap between the possible, the probability of the possible, and the probability of actual entrenchment into lexicon.

A Lexical borrowings found in *Le Monde* (T1)

Borrowings are given as they appear in the T1 corpus. A dash indicates that the type was excluded from the *Le Figaro* (T2) study for reasons discussed in Section 3.2. The sense *offshore_1* pertains to drilling for oil at sea, while *offshore_2* qualifies business conducted abroad. One occurrence of *offshore* was excluded because its sense could not be determined.

Borrowing	Frequency		Borrowing	Frequency	
	T1	T2		T1	T2
A contrario	1	667	inquilinaje	2	0
Canada Dry	1	32	investment banks	1	4
Errare humanum est , perseverare diabolicum	1	1	joint - venture	17	400
Gross Up	1	1	joint - ventures	2	221
Just do it	1	21	junk bonds	2	127
Karenztag	2	0	khazjajstvo	1	0
Lander	6	1147	kids	2	15
Last but not least	1	178	kippa	1	183
Mismanagement	1	4	know how	1	15
Nehruian Socialism	1	0	kollektivnoe	1	0
TIERRA Y LIBERTAD	1	0	land	1	–
afrikaans	1	17	latifundio	3	1
alya	1	18	lean production	1	0
an elastic currency	1	0	lease - back	2	10
apparatchik	1	289	leitender Angestellter	1	0
apple - pie	1	3	lobbying	3	865
bag - ladies	1	0	look	2	1387
basics	1	22	lottizzazione	1	0
because	1	5	male oscuro	1	0
big Three	1	101	markka	1	13

big bang	3	489	medium - term notes	2	0
board	3	184	merchant banks	1	4
boat - people	1	124	minifundio	3	0
brain storming	1	23	money funds	1	0
brocca	1	0	names	1	14
bush	1	–	news	2	120
campo	1	7	next jump	1	0
cash	1	2120	next root	1	0
cash flow	2	711	next wave	1	0
cassa integrazione	2	2	nomenklatura	4	190
casualwear	1	6	offshore	1	–
chapka	1	55	offshore_1	5	300
check - up	1	59	offshore_2	1	274
chehita	1	0	old lady	1	4
citizen' s charter	1	0	open market	1	4
classless society	1	0	outsourcing	1	98
come - back	1	333	pack	2	–
credit crunch	1	23	parity cracking	2	0
cross borders	1	0	perestroika	8	210
debt deflation	1	6	ph.D	1	11
deficiency payments	1	1	popiwek	1	0
deregulation	1	–	prime rate	1	6
deutschemark	25	154	prorata temporis	1	61
deutschemarks	24	223	res nullius	1	2
discount	1	305	roadbook	1	1
downgrading	1	0	running	1	53
dynamic random access memory	1	3	savings and loans	1	6
ejideros	2	0	scala mobile	4	3
ejido	9	0	shipping	2	2
ejidos	8	1	short	1	–
establishment	2	500	show - room	1	165
estancias	1	9	sommerso	1	0
ex- joint - venture	2	0	stand - by	3	123
fazendas	1	5	stop loss	1	6
fincas	1	12	struggle for life	1	8
flash	3	33	success story	3	382
flint glass	1	0	sustainable	1	1
float glass	1	3	swaps	1	71
french doctors	1	25	sweats	1	4
french mafia	2	0	taref	1	0
french travel way of life	1	0	teddy	1	18
geschäftsfürhung	1	0	terratenientes	2	0
glasnost	2	82	the	1	0
government	1	0	to regulate	1	0

half baked	1	0	top - down	1	25
hedge funds	3	368	trade unions	1	2
high - tech	2	2056	training groups	1	0
homeless	1	6	tripalium	1	4
huasipongo	2	0	under - class	1	1
industrial design	1	0	welfare state	2	9

References

Abeillé, A., L. Clément, & F. Toussenel. 2003. Building a treebank for French. In *Treebanks: Building and Using Parsed Corpora*, 165–188. Kluwer Academic Publishers.

Agresti, A. 2002. *Categorical Data Analysis*. New York: Wiley, 2nd edition.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19.716–723.

Aronoff, Mark. 1976. *Word Formation in Generative Grammar*. Cambridge, Mass.: MIT Press.

Baayen, R.H. 1992. Quantitative aspects of morphological productivity. In *Yearbook of Morphology 1991*, ed. by G. E. Booij & J. van Marle, 109–149. Dordrecht: Kluwer Academic Publishers.

——, D. J. Davidson, & D.M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* p. in press.

——, & R. Lieber. 1997. Word frequency distributions and lexical semantics. *Computers and the Humanities* 30.281–291.

——, & A. Renouf. 1996. Chronicling the Times: Productive Lexical Innovations in an English Newspaper. *Language* 72.69–96.

Bauer, L. 1983. *English Word Formation*. Cambridge: CUP.

Bybee, J., & P. Hopper (eds.) 2001. *Frequency and the emergence of linguistic structure*. Amsterdam: Benjamins.

Dendien, J., & J.-M. Pierrel. 2003. Le trésor de la Langue Française informatisé: un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement automatique des langues* 44.11–37.

Erling, E., 2004. *Globalization, English, and the German University Classroom*. University of Edinburgh dissertation.

Etiemble, R. 1964. *Parlez-vous franglais ?*. Paris: Gallimard.

Flaitz, J. 1988. *The Ideology of English: French Perceptions of English as a World Language*. Berlin: Mouton de Gruyter.

Fuchs, C.. 1996. *Les ambiguïtés du français*. Paris: Editions Ophrys.

Gries, S.T. 2004. Shouldn't it be breakfunch? A quantitative analysis of blend structure in English. *Linguistics* 42.639–667.

—— 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13.

Hagège, C. 2006. *Combat pour le français au nom de la diversité des langues*. Paris: Odile Jacob.

Halle, M., & A. Marantz. 1993. Distributed morphology and the pieces of inflection. In *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*, ed. by K. Hale & S. J. Keyser, volume 24 of *Current Studies in Linguistics*, 111–176. Cambridge, Mass: MIT Press.

Harrell, F.E. 2001. *Regression modeling strategies*. Berlin: Springer.

Howes, D., & R. L. Solomon. 1951. Visual duration threshold as a function of word probability. *Journal of Experimental Psychology* 41.401–410.

Klein, D., & G. L. Murphy. 2001. The Representation of Polysemous Words. *Journal of Memory and Language* 45.259–282.

Metcalf, A. 2004. *Predicting New Words: The Secrets of Their Success*. Boston: Houghton Mifflin Harcourt.

New, B., L. Ferrand, C. Pallier, & M. Brysbaert. 2006. Re-examining word length effects in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review* 13.45–52.

Oldfield, R. C., & A. Wingfield. 1965. Response latencies in naming objects. *Quarterly Journal of Experimental Psychology* 17.273–281.

Parks, R., J. Ray, & S. Bland, 1998. *Wordsmyth English dictionary-thesaurus*. University of Chicago. `http://www.wordsmyth.net/Chicago`.

Pergnier, M. 1989. *Les anglicismes: danger ou enrichissement pour la langue française ?*. Paris: Presses universitaires de France.

Picone, M.D. 1996. *Anglicisms, Neologisms, and Dynamic French*. Amsterdam: John Benjamins.

Piercey, C.D., & S. Joordens. 2000. Turning an advantage into a disadvantage: Ambiguity effects in lexical decision versus reading tasks. *Memory and Cognition* 28.657–666.

Pinker, S. 2007. *The Stuff of Thought: Language as a Window into Human Nature*. New York: Viking Adult.

Rey-Debove, J. 1987. Effet des anglicismes lexicaux sur le système du français. *Cahiers de lexicologie* 51.257–265.

Rodd, J.M., M.G. Gaskell, & W.D. Marslen-Wilson. 2004. Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science* 28.89–104.

——, & W.D. Marslen-Wilson. 2002. Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language* 46.245–266.

Selkirk, E. 1982. *The Syntax of Words*. Cambridge: The MIT Press.

Steels, L., F. Kaplan, A. McIntyre, & J. Van Looveren. 2002. Crucial factors in the origins of word-meaning. In *The Transition to Language*, ed. by Alison Wray, chapter 12. Oxford: Oxford University Press.

Strobl, C., T. Hothorn, & A. Zeileis. 2001. Party On! a New, Conditional Variable Importance Measure for Random Forests Available in the `party` Package. Technical Report 50, Institut für Statistik, Ludwig-Maximilians-Universität München.

Thogmartin, C. 1988. The public impact of terminological planning in France. *The French Review* 64.1000–1006.

Thomason, S.G., & T. Kaufman. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.

Ussishkin, A. 2005. A Fixed Prosodic Theory of Nonconcatenative Templatic Morphology. *Natural Language & Linguistic Theory* 23.169–218.