# Frequency in lexical processing

R. Harald Baayen, Petar Milin, and Michael Ramscar

Eberhard Karls University, Tübingen, Germany

## Abstract

This study is a critical review of the role of frequency of occurrence in lexical processing, in the context of a large set of collinear predictors including not only frequencies collected from different sources, but also a wide range of other lexical properties such as length, neighborhood density, measures of valence, arousal, and dominance, semantic diversity, dispersion, age of acquisition, and measures grounded in discrimination learning. We show that age of acquisition ratings and subtitle frequencies constitute (reconstructed) genres that favor frequent use for very different subsets of words. As a consequence of the very different ways in which collinear variables profile as a function of genre, the fit between these variables and measures of lexical processing depends on both genre and task. The methodological implication of these results is that when evaluating effects of lexical predictors on processing, it is advisable to carefully consider what genres were used to obtain these predictors, and to consider the system of predictors and potential conditional independencies using graphical modeling.

Frequency of occurrence is perhaps the strongest and most-studied predictor of lexical processing. Counts of occurrences of words (Rayner and Duffy, 1986; Gardner, Rothkopf, Lapan and Lafferty, 1987; Glanzer and Bowles, 1976; Grainger, 1990; Griffin and Bock, 1998; Jescheniak and Levelt, 1994; McRae, Jared and Seidenberg, 1990; Meunier and Segui, 1999; Scarborough, Cortese and Scarborough, 1977; Stemberger and MacWhinney, 1986; Wingfield, 1968; Baayen, Wurm and Aycock, 2007, 2010; Halgren et al., 2002; Young and Rugg, 1992), of syllables (Carreiras, Alvarez and de Vega, 1993; Cholin, Schiller and Levelt, 2004; Barber, Vergara and Carreiras, 2004), and word n-grams (Tremblay and Baayen, 2010; Tremblay, Derwing, Libben and Westbury, 2011; Bannard and Matthews, 2008; Arnon and Snider, 2010; Shaoul, Westbury and Baayen, 2013; Ramscar, Hendrix, Shaoul, Milin and Baayen, 2014; Shaoul, Baayen and Westbury, 2015) have been shown to correlate well with chronometric measures such as response latencies, with many aspects of the eye-movement record, and with the brain's electrophysiological response to lexical stimuli. Frequency of occurrence is also predictive for many aspects of lexical form, including acoustic duration, length in phones or letters, tone, and pitch (Zipf, 1929; Gahl, 2008; Pluymaekers, Ernestus and Baayen, 2005; Wright, 1979; Tomaschek, Wieling, Arnold and Baayen, 2013, 2014; Zhao and Jurafsky, 2009; Koesling, Kunter, Baayen and Plag, 2012; Gahl, Yao and Johnson, 2012; Arnon and Priva, 2013).

Although it appears to represent a deceptively simple concept, frequency of occurrence in language, and in the mental lexicon in particular, actually turns out to be a remarkably complex construct that comprises a large set of highly collinear lexical random variables. The goal of the present study is to clarify the place of frequency of occurrence in this complex system, paying attention in particular to the relationship between frequency and dispersion, register, age of acquisition, and response times in visual lexical decision tasks.

In what follows, we first provide a critical assessment of the issues, and then outline a novel way for understanding how frequency effects come about in lexical processing. We next use graphical modeling to present an analysis of the full collinear system of factors influencing "frequency" , and conclude with some practical considerations as to how the surprisingly complex concept of lexical frequency might be best approached in studies of language and language processing. We hope the perspective on frequency of occurrence and its consequences for lexical processing in healthy brains will help inform investigations of the breakdown in aphasia of lexical processing under physiological insult.

We begin with considering the thorny question of what it is *exactly* that gets counted when matters of frequency of lexical occurrence are assessed.

## 1 The units of counting

In measuring lexical frequency, we immediately encounter a question: *frequency of what?* What *exactly* are the lexical "units" that we are supposed to count? Among the first groups of scholars to ever systematically address this question were the Masoretes in the 6th to 10th centuries, who meticulously counted words, letters, and certain collocations in the Hebrew scriptures for the purposes of standardization and ensuring quality control over texts and their dissemination. To do so, they turned to the technology of writing to determine what got counted, establishing a textual hegemony over lexical measurement that has endured across the centuries.

However, across languages, textual and orthographic practices vary enormously in the way that they discretize the continuous and linguistically more primary medium of speech, and this means in turn that they offer up very different basic metrics when it comes to measurement. English conventionally uses space characters to separate words, and for those whose first language is English and whose first training in literacy is in English, the word is a natural, self-evident, given. Yet by

contrast, the alphabetic writing system of Vietnamese uses space characters to separate syllables rather than words. Chinese characters typically correspond to syllables and, as in Vietnamese, these syllables simultaneously have morphemic status. Given the very different textual conventions of Vietnamese, linguists have coined the term *syllabeme* to describe the orthographic units that they are presented with (Nguyen, 2011; Pham, 2014). Meanwhile, the hangul alphabet of Korean groups letters into syllabic configurations, which in turn group together to form words. As these comparisons hopefully make clear, the basic "lexical units" that are delimited by orthographical conventions turn out to be remarkably language-specific.

Accordingly, in the present digital world, orthographic conventions continue to determine to perhaps a surprising degree what is amenable to (computerized) counting. Consider the writing conventions of the Germanic languages English, German, and Dutch. English splits many of its onomasiological units into multiple orthographic words, both in compounds (*ring binder*, *engagement ring*), verb-particle combinations (*ring up*, 'to telephone'; *ring out*, 'to sound the bells that announce weddings etc.), and in idioms (*run rings around someone*, in the sense of obviously outperforming someone). By contrast, in German and Dutch, compounds are always written without intervening spaces, and particle-verb combinations are written as single words whenever the particle immediately precedes the verb in the sentence (e.g., Dutch *appeltaart*, 'apple pie'; German *anrufen*, 'to ring up' versus *ruft an*, 'rings up'). As in English, idioms are always spaced. As a consequence of these different writing conventions, counts for *ring* in English will include tokens of the letter sequence `ring` as a constituent in compounds and particle verbs, and as part of idioms. Counts for the corresponding cognates in German and Dutch will include idioms (albeit typically very different idioms) and occurrences of `ring` in particle verbs in those constructions where verb and particle are separated, but not occurrences of `ring` as a constituent in compounds. Cumulation of frequencies across distinct onomasiological units is particularly widespread in Hebrew, because many vowels are not actually specified in common orthographic practice. As a consequence, homography in Hebrew is rampant.

Computational linguists have, to date, been unable to develop algorithms that reliably identify onomasiological units in English (compounds, verb-particle combinations, or idioms) written with intervening space characters. Whether one consults the CELEX lexical database (Baayen, Piepenbrock and Gulikers, 1995), the British National Corpus (Burnard, 1995), the Corpus of Contemporary American English (Davies, 2010), or corpora constructed from film subtitles (Brysbaert and New, 2009), it is invariably the case that what ultimately gets counted is determined in large part by whatever the strings of letters that are separated by spaces turn out to be (perhaps enriched with tags for part of speech, etc.). As a recent example, van Heuven, Mandera, Keuleers and Brysbaert (2014) decided to remove all hyphens in a corpus of television subtitles, and motivated this strategy with the observation that the resulting frequency counts were better able to predict reaction times. Whereas this strategy may perhaps be a reasonable choice for adjective-noun combinations (as in *a life-saving drug*), it has as adverse side-effect that now not only spaced compounds (*apple pie*) are invisible to the researcher, but also those compounds which in the original text are identifiable as lexicalized onomasiological units thanks to the hyphen (Kuperman and Bertram, 2013).[1]

The fact that differing orthographic conventions result in substantial between-language variability in what is counted is not the only problem one encounters in measuring "lexical frequency". Languages also vary enormously in their structural properties, and this contributes a second source of cross-linguistic variation when it comes to counting "lexical events". Words in polysynthetic languages can express what in English would require multi-word phrases (e.g., Greenlandic Eskimo

[1] A further disadvantage of this strategy is that it comes with the danger of circularity: Frequency counts collected to predict lexical processing are themselves based on decisions about data preprocessing that are informed by how well candidate counts predict lexical processing.

*tikitnikuusimavoq*, 'apparently, she had arrived'). Languages with rich verbal or nominal inflectional paradigms such as Italian and Estonian likewise usually express in one form that which English ordinarily discretizes into multiple pronouns, auxiliaries and prepositions (Italian *finivamo*, 'we finished'; Estonian *kivisse*, 'into the stone'). A straightforward consequence of this is that when languages have rich inflectional morphology, frequency counts tend to be characterized by substantially greater word form type frequency and much lower token frequency, as compared to languages with sparser inflectional morphology such as English and Dutch, or Chinese and Vietnamese.

A third factor determining what is counted is the overwhelming culture of literacy in which research on lexical processing is carried out. Although frequency counts are based on orthographic conventions, these conventions are in many ways far removed from the actual forms that are prevalent in the spoken language. The printed word suggests an invariance that is absent in speech. The spoken word is informative about a speaker's sex, age, social background, emotional state of mind, and a wealth of other information that is totally absent in print. Examination of corpora of spontaneous speech has revealed that many words are realized with shortened forms, with segments or even entire syllables missing (Johnson, 2004; Keune, Ernestus, Van Hout and Baayen, 2005; Pluymaekers et al., 2005; Ernestus, Baayen and Schreuder, 2002). For instance, English *yesterday* can be realized as /jɛʃeɪ/, and Dutch *natuurlijk*, with as canonical pronunciation /nʌtyrlək/, appears in many different shortenings, including /tyrlək/, /tyk/, and /tək/. Johnson (2004) reports for English a 5% deletion rate of syllables, a 25% deletion rate of segments in content words, and deletion rates up to 40% for function words. In addition, many words are realized with other segments than those given by their canonical form.

The actual complexities of speech raise questions regarding the determination of similarity and difference — i.e whether two items represent two types or two tokens of a type — that are obscured by the arbitrary nature of orthographic conventions. Although the standard classification of English words such as *time* and *thyme* as homonyms suggests they share the same invariant phonic form, it has been shown that their acoustic realizations are statistically distinct (Gahl, 2008). Thus, the speech signal is much more varied and distinctive than orthographic conventions or phonological transcriptions of canonical forms suggest. As a consequence, counts based on written texts will often not reflect form differentiation characteristic of spoken language.

Simultaneously, the invariability of words suggested by printed text contributes to a pre-scientific way of thinking about word meanings, where words are typically taken to express one meaning. Homonyms, in other words, are typically viewed as exceptions rather than the norm. However, many common function words such as *in, but, we, not, one, some, would, no*, and *our*) have homophones (*inn, butt, wee, knot, won, sum, wood, know*, and *hour*), indicating that in speech they are more similar to each other than one might assume based on their spelling.

Further, the number of a word's meanings and senses increases with frequency of occurrence (Köhler, 1986; Baayen and Moscoso del Prado Martín, 2005). A fairly high-frequency content word such as English *ring* comes with a bewildering number of meanings and senses, including 'a circular ornamental band of metal worn on the finger', 'an inclosed area for a sports contest', 'a group of persons cooperating for unethical or illicit purposes', 'to encircle', 'to give forth a clear resonant sound', 'a telephone call', and 'the impression created by a statement' (as in 'her story had a ring of truth'). As a consequence, counts based on English words aggregate over many meanings and senses that in other languages may well be expressed by a variety of etymologically unrelated words. Since such different meanings are typically self-evident when words appear in context, counts of space-separated letter strings are decontextualized counts.

In summary, what is (typically) counted is what happens to be written in a given language with distinct orthographic forms. These orthographic forms may be quite different from the forms realized in speech. Especially for higher-frequency words, the forms counted are onomasiologically

heterogeneous. The morphological characteristics of a language furthermore determine the extent to which even for one meaning or sense, counts are fractionated across inflectional variants.

## 2 Corpora and constraints on counting

### 2.1 The corpus as a mirror of collective experience

Early counts of word occurrences were carried out by hand, either in an educational context (see, e.g., Thorndike & Lorge, 1944), or from a statistical interest (Zipf, 1935; Yule, 1944). The earliest digital corpus was compiled in the early sixties at Brown university, and comprised one million work tokens. Word frequency counts based on this corpus were distributed in book form (Kučera and Francis, 1967). Although an impressive achievement, both with respect to the careful sampling of textual materials and given the limited computational resources of the time, the sample size of the Brown corpus is, in retrospect, far too small to afford sufficient precision for research on language processing.

Given the historical limitations of resources such as the Brown corpus, Gernsbacher (1984) suggested that subjective frequency estimates collected from experimental subjects might be used instead. However, it turns out that when subjects are asked to rate how frequent a word is, they are unable to provide estimates of pure frequency. Rather, analyses have revealed that their judgments are contaminated by the many other lexical dimensions that correlate with frequency of occurrence, such as dimensions of emotionality (Baayen, Feldman and Schreuder, 2006; Westbury, 2014).

Turning to the present, much larger corpora are now available for English, such as the British National Corpus (BNC, 100 million become words, Burnard, 1995), the Corpus of Contemporary American English (COCA, 450 million words Davies, 2010), corpora harvested from the web for several languages with more than 1 billion words each (Baroni, Bernardini, Ferraresi and Zanchetta, 2009), and the frequency lists published by Google, which are based on a 1 trillion word sample from the web (Brants and Franz, 2006).

Speech corpora are, however, less common, and typically much smaller. The British National Corpus comprises 10 million words of speech, of which 5 million were sampled from free, unscripted conversational speech. For Dutch, a spoken corpus of similar size is available as well (Oostdijk, 2002). For American English, the Buckeye Corpus (Pitt, Johnson, Hume, Kiesling and Raymond, 2005) is an important source of information on the acoustic properties of conversational speech, thanks to its excellent phonetic mark-up. The ONZE corpus (Gordon, Maclagan and Hay, 2007) is a rich speech corpus of New Zealand English, and famous for the unique perspective it offers on the phonetics of language change.

The construction of speech corpora is very labor intensive and extremely expensive compared to building corpora of written language. In order to perhaps better approximate everyday spoken language, corpora consisting of film subtitles, which are straightforward to extract from existing resources on the web, have recently been compiled (New, Brysbaert, Veronis and Pallier, 2007; Brysbaert and New, 2009; Brysbaert, Keuleers and New, 2011, 2015). Due to copyright restrictions, these corpora are not generally available, but word frequencies and related statistics are copyright-free, and can be found, for instance, at http://crr.ugent.be.

An assumption that lies behind the use of corpora in much psycholinguistic work is that a suitably representative corpus of, say English, can serve to represent (or control for) subjects prior lexical experience in accounting for various aspects of linguistic behavior. There is, however, reason to believe that the nature (and in particular, the statistical properties) of linguistic experience serves to undermine this assumption. For example, frequencies of occurrence vary across regional varieties, as attested for English by a family of corpora, that, following the model of the Brown

corpus, have been constructed for British English, Australian English, Indian English, Canadian English, and New Zealand English (Xiao, 2008). Furthermore, frequency counts vary as well with register and text type (Biber, 1988, 1989), and how frequently individual writers use their words provides a statistical fingerprint of their authorial hand (Burrows, 1987, 1992; Halteren, Baayen, Tweedie, Haverkort and Neijt, 2005).

The diversity of lexical usage and experience indicate that in using frequency counts for the study of specific aspects of lexical processing, it is important to consider the communicative goals of the texts sampled by a given corpus, and the specific demands imposed by a given task probing aspects of lexical processing. To illustrate the impact of these factors on the way that this complicates the interpretation of "frequency effects", we consider them in relation to frequency counts based on corpora of film subtitles, which have recently become popular as measures of lexical frequency.

Film subtitle frequency counts have have been found to provide improved predictivity for reaction times compared to standard text-based frequency counts. Brysbaert and New (2009) take this to indicate that subtitles can thus be considered to better approximate language as it is used on a daily basis. Indeed, the impression one gains from this literature is that, for the assessment of language processing in general, subtitle corpora can be taken as the source for normative measures of lexical frequency.

Yet, as the reasons we described above indicate, from a linguistic perspective, this state of affairs is puzzling. First, why should one particular register of language use have such a pre-eminent status for language processing in general? Wouldn't one expect that when reading a novel, the frequencies (as well as co-occurrence frequencies, probabilities and surprisals) particular to novels (as a genre) be more precise as predictor of readers' expectations? Second, why, of all registers in modern language communities, should the register of film subtitles specifically have proved to be such a pre-eminently reliable predictor of lexical processing?

This latter finding is especially surprising because film subtitles are twice removed from spontaneous conversations in day-to-day communication. The conversations in films are scripted, and on top of this, the actual subtitles shown on screen tend to reflect the gist of what is being said, rather than reporting the utterances in the film verbatim, as a result of the constraints imposed by the medium (e.g., having to avoid multi-line subtitles that may be too long to read in the available time).

So why might frequencies culled from subtitles prove to be so successful at predicting reaction times in the lexical decision and word naming tasks? One important part of the answer is offered by Heister and Kliegl (2012), who report that for German, frequencies extracted from a tabloid newspaper (Bild Zeitung) have similar predictive value as frequencies from a German subtitle corpus. They also obtained similar results for frequencies collected from a 1.2 billion word `deWaC` web corpus (Baroni et al., 2009). Notably, the performance of both subtitle and tabloid frequencies was notably better for words with positive or negative valence, prompting the authors to suggest that it is emotional language rather than the approximation of spoken language that lies at the heart of the success of subtitle frequencies. (The study also showed that subtitles tend to repeat words more often, and to make use of shorter words.)

In the light of these German findings, we examined in detail an English data set which consists of 4440 words that occur in the child-directed speech of the English subset of the CHILDES database (MacWhinney, 2000), and for which emotion ratings (Warriner, Kuperman and Brysbaert, 2013), as well as subtitle frequencies and reaction times from the British Lexicon Project (Keuleers, Lacey, Rastle and Brysbaert, 2012) are available.

To this data set, we added written and spoken frequencies from the British National Corpus, using for the spoken frequencies the demographic subcorpus. This subcorpus provides transcripts of recordings made of speakers of different ages, socio-economic status, and geographic location

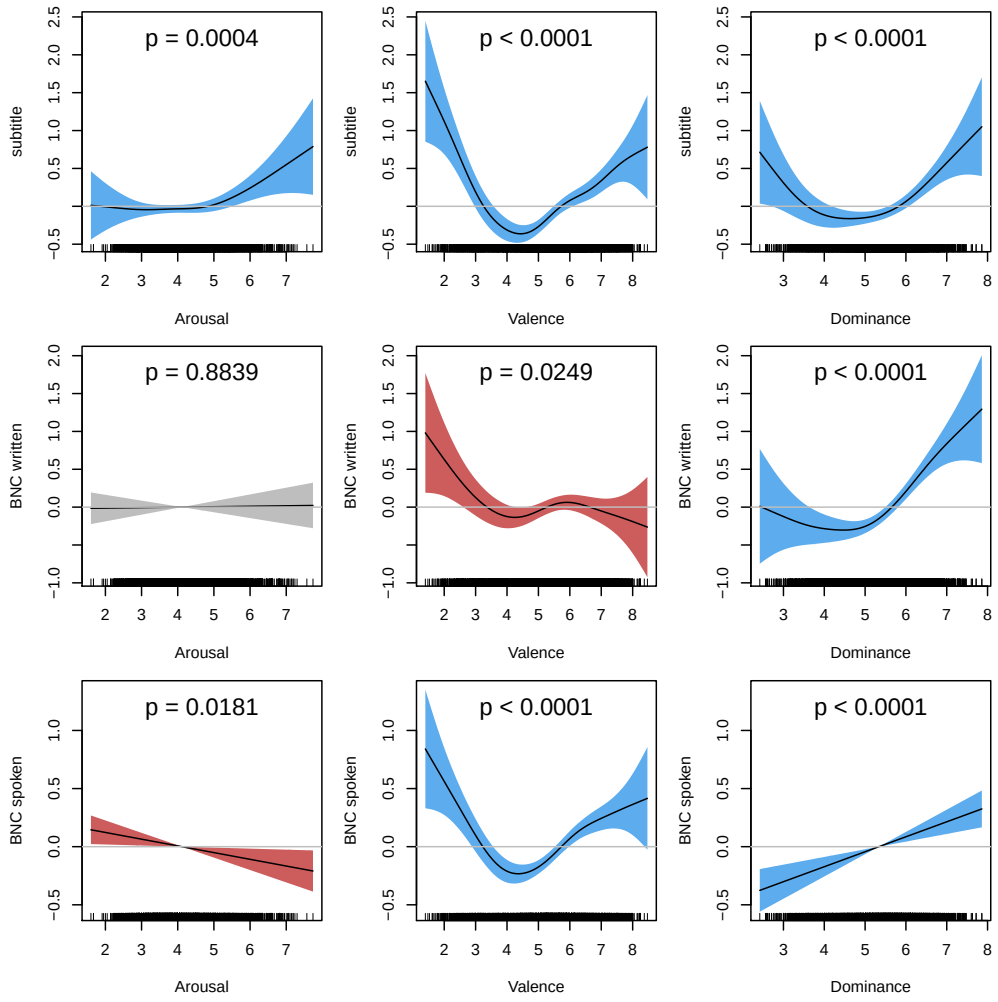


Figure 1: Partial effects for arousal, valence and dominance as predictors of log subtitle frequency (top panels) and log BNC written frequency (center panels), and log BNC spoken frequency (bottom panels). Blue: well-supported effects, red: marginal effects, grey: no effect.

in England. Each informant was supplied with a small Walkman and a microphone. They were requested to record all speech, both their own and the speech of others, over a period of one week. These recordings contain highly natural speech which comes as close as possible to normal everyday language. With 5 million word tokens, this corpus of spoken English is large enough to allow systematic comparisons between both written English and subtitle English to be made.

Figure 1 presents the partial effects of arousal, valence, and dominance as predictors of log subtitle frequency[2] (top row), of log BNC written frequency (second row), and of log BNC spoken frequency (bottom row), obtained with thin plate regression splines[3] as available in the `mgcv` package for `R` for generalized additive models (Wood, 2006; Baayen, 2013).[4] Analyses were carried out with

[2] We backed off from zero by adding 1 to the corpus frequency before taking the (natural) logarithm.

[3] A thin plate regression spline approximates a wiggly curve as a weighted sum of mathematically regular curves (named basis functions), with a penalty on wiggliness. The estimation algorithms make sure a good balance is found between fidelity to the data and model simplicity.

[4] Data sets and analyses reported in this manuscript are available in the Mind Research Repository at http://openscience.uni-leipzig.de/index.php/mr2.

both the subtitles available from http://crr.ugent.be and, to keep corpus size comparable, a 5 million word subtitle corpus sampled from an 1100 million word subtitle corpus we assembled ourselves. (It is important to note that while the following analyses use this 5 million word subtitle corpus, similar results were obtained in a further set of analyses employing the subtitle frequencies given on the Ghent website. For the present data set, the correlation of our counts and those from Ghent is very high, 0.974, indicating that we successfully replicate the Ghent subtitle frequency estimates.)

A comparison of the leftmost panels in each row reveals that higher frequency words in subtitles tend to be high arousal words, whereas in actual British conversation, higher-frequency words have arousal values that decrease with frequency. Further, arousal is not predictive at all for written English (first panel, second row). Taken together, these findings indicate that in normal English conversation, highly arousing words are used sparingly, whereas perhaps unsurprisingly given the dramatic nature of film, these words enjoy *far* more popularity in subtitles.

Next, with respect to valence (the second column of panels, which contrast unhappy and unpleasant words with happy and pleasant words), low valence predicts high frequency of use across subtitle, written, and spoken English. Further, the written corpus is unique in that a high valence does not predict greater frequency of use. With this in mind, it is noteworthy that the effect size of valence is much larger in subtitles (where the mean varies from 1.5 to -0.25 to 0.8) than it is in conversational English (where the mean varies from 0.8 to -0.2 to 0.4). In other words, in comparison to the other corpora, it would appear that film subtitles overuse happy and sad words.

The third column of panels gauges the extent to which a word is associated with weakness and submissiveness versus strength and dominance (e.g., *doomed* versus *won*). As can be seen, subtitle English largely resembles written English when it comes to dominance, with the main difference between the two being that in the latter, words with lower dominance values are used less frequently. In true conversational English, by contrast, the effect of dominance is linear, with a positive slope, indicating that lower dominance and less intensive use go hand in hand. For this register, the effect size of dominance is also slightly reduced compared to subtitle English.

Figure 2 plots word length, orthographic neighborhood density and the the number of meanings/senses per word (gauged by means of the number of synsets in WordNet, Miller, 1990, in which the word appears) in both the subtitle frequency corpus (top panels) and the spoken BNC corpus (bottom panels). As can be seen, normal spoken English differs from subtitle English on all of these measures, both qualitatively and quantitatively in the case of word length, and quantitatively for the synset and neighbor counts, with somewhat larger effect sizes for the subtitles. Further model comparisons (not shown) support the pairwise differences visible in Figure 2. Thus, subtitle English appears to make use of a more "amplified" register. As magnitudes of these effect are greater for the subtitle frequencies than the spoken BNC frequencies, it appears that subtitles make more intensive use of words with many meanings, while avoiding the use of words with many neighbors as well as longer words. Indeed, in this last respect it appears that the constraint of having to keep film subtitles short gives rise to a very important difference with more usual conversational English.

To summarize: in our comparison of English subtitles to English spoken and text corpora, we observed a pattern of results that is highly consistent with what Heister and Kliegl (2012) found in German. Compared to normal day-to-day conversational English, subtitles are characterized by more intense use of high-arousal words, and of words with more extreme values of valence and dominance. This makes perfect sense for a genre that ultimately reflects the economic reality of films: to provide its audience with emotionally rich experiences, along with other related constraints, such as the fact that subtitles need to be both quick and easy to read. Given these constraints, the fact that subtitle writers tend to a more amplified register (using shorter words, with more meanings or senses, and fewer orthographic competitors) seems to be a natural and highly adaptive response.

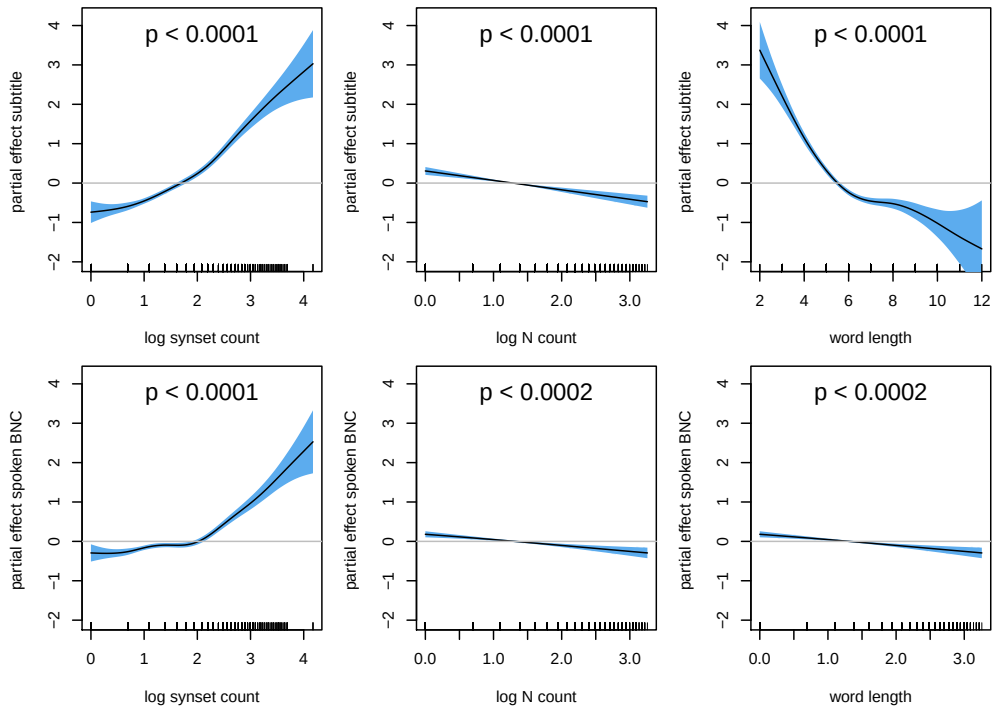


Figure 2: Partial effects for number of synsets, number of orthographic neighbors, and word length (in letters) as predictors of log subtitle frequency (top panels) and log BNC spoken frequency (bottom panels).

## 2.2 The corpus as a predictor of processing

The subtle ways in which lexical distributional properties vary across text types has far reaching consequences for the statistical analysis of measures of lexical processing. Figure 3 presents the effects on log RT of number of senses (operationalized as above), number of orthographic neighbors, word length, frequency, arousal, valence, and dominance. (In this analysis, as in all analyses to follow, all of the predictors were scaled in order to ensure optimal parameter estimation.) The upper panels pertain to a model (AIC: -10919) in which subtitle frequency was included. The lower panels present the corresponding model in which subtitle frequency is replaced by BNC spoken frequency (AIC: -10445). A comparison of the two models, reveals that the former has the superior fit, along with two further noteworthy facts:

1. In the model employing subtitle frequencies, lexical frequency is a *stronger* predictor of response latencies than is the case for the model in which lexical frequencies were taken from the spoken BNC corpus. As can be seen in the fourth panel on the second row, the frequency effect levels off quickly for the higher BNC frequencies.

2. The effects of all of the other six predictors are *weaker* for the model that employed subtitle frequencies, and stronger for the model that employed spoken BNC frequencies.

This difference can be quantified using Akaike's information criterium (AIC; a standard metric for evaluating the quality of statistical models while controlling for the inevitable trade-off between complexity and goodness of fit). Table 1 lists the reduction in AIC obtained by first adding to a baseline model with frequency as only predictor the three lexical predictors — number of senses

(synsets), number of neighbors and length — and in a second step, the effects of adding the three emotion predictors, arousal, valence and dominance. As is clear from Table 1, the reductions in AIC are substantially larger when these data are modeled using BNC spoken frequencies than when the same frequencies are derived from a subtitle corpus, a finding that makes sense given that, as we showed above, the BNC spoken frequencies are less well-predicted by these six measures.

| | lexical predictors | emotion predictors |
|---|---|---|
| subtitle frequency | 195.45 | 134.75 |
| BNC spoken frequency | 356.18 | 222.06 |

Table 1: The amount by which Akaike's information criterium (AIC) is reduced when lexical variables (left) and emotion variables (right) are added to a model with subtitle frequency and BNC spoken frequency.

These findings strongly indicate that when it comes to modeling tasks such as visual lexical decision and word naming, subtitle frequencies do not provide excellent fits because they provide a more accurate representation of the frequency information underlying participants responses. Rather, it seems that subtitle writers use short, simple, and emotionally laden words more frequently, and this produces in a highly readable, emotionally charged register that is optimized for its function: rapid visual uptake of lexical information in a medium (film) where the predominant visual emphasis is quite definitely *not* textual. Rapid visual uptake is, of course, exactly what is required in speeded lexical decision and word naming tasks, when words are presented in isolation, bereft of the rich contexts in which they occur in normal language use. And this indicates that frequencies taken from subtitle corpora provide excellent fits for this kind of data *not* because they capture the *frequency* information that drives participants' responses in them, but rather because, as a register, subtitles serve to strongly confound frequency with a number of other variables that also contribute to faster of slower lexical responses.

Further, if our explanation of the superiority of subtitle frequencies for lexical decision and naming is correct (i.e., if the subtitle register confounds various factors that optimizes its fits for these specific tasks), it leads to a clear prediction: If we consider lexical processing in a task and register that we would not expect to be attuned to the specific constraints that shape subtitles, for example reading English novels (where the predominant visual emphasis quite definitely *is* textual), and if we exchange isolated word presentation with reading in normal discourse context, and replace lexical decision by an eye-tracking measure such as first fixation duration, then we should expect that subtitle frequencies might no longer be the best predictor of behavior. Indeed, we might even expect subtitles to provide inferior fits as compared to frequency counts based on normal written language use.

To test this prediction, we examined a set of eye-movement data collected while a total of four participants read through the subcorpus of fiction in the Brown corpus (Hendrix, 2015), re-analyzing a set of 316 compounds types in the subcorpus that were fixated only once in reading (the reading pattern for 60% of the tokens). In an earlier analysis of this set, Hendrix observed that, in interaction with the LSA similarity (Landauer and Dumais, 1997) of the compound and its first constituent, the frequency of the compound taken from the British National Corpus was a good predictor of fixation durations in reading. When we tested to see what would happen when Hendrix's original analysis was repreated using frequencies taken from our 1100 million word subtitle corpus, we found that exchanging the BNC frequencies for subtitle frequencies caused the goodness of fit of the model to decrease (the AIC score went up by 7 units). Or in other words, once tasks and measures that are particularly suited to the subtitle register (speeded lexical decision making in response to isolated words) are replaced by response measures (eye-movements) and tasks sensitive to the way that

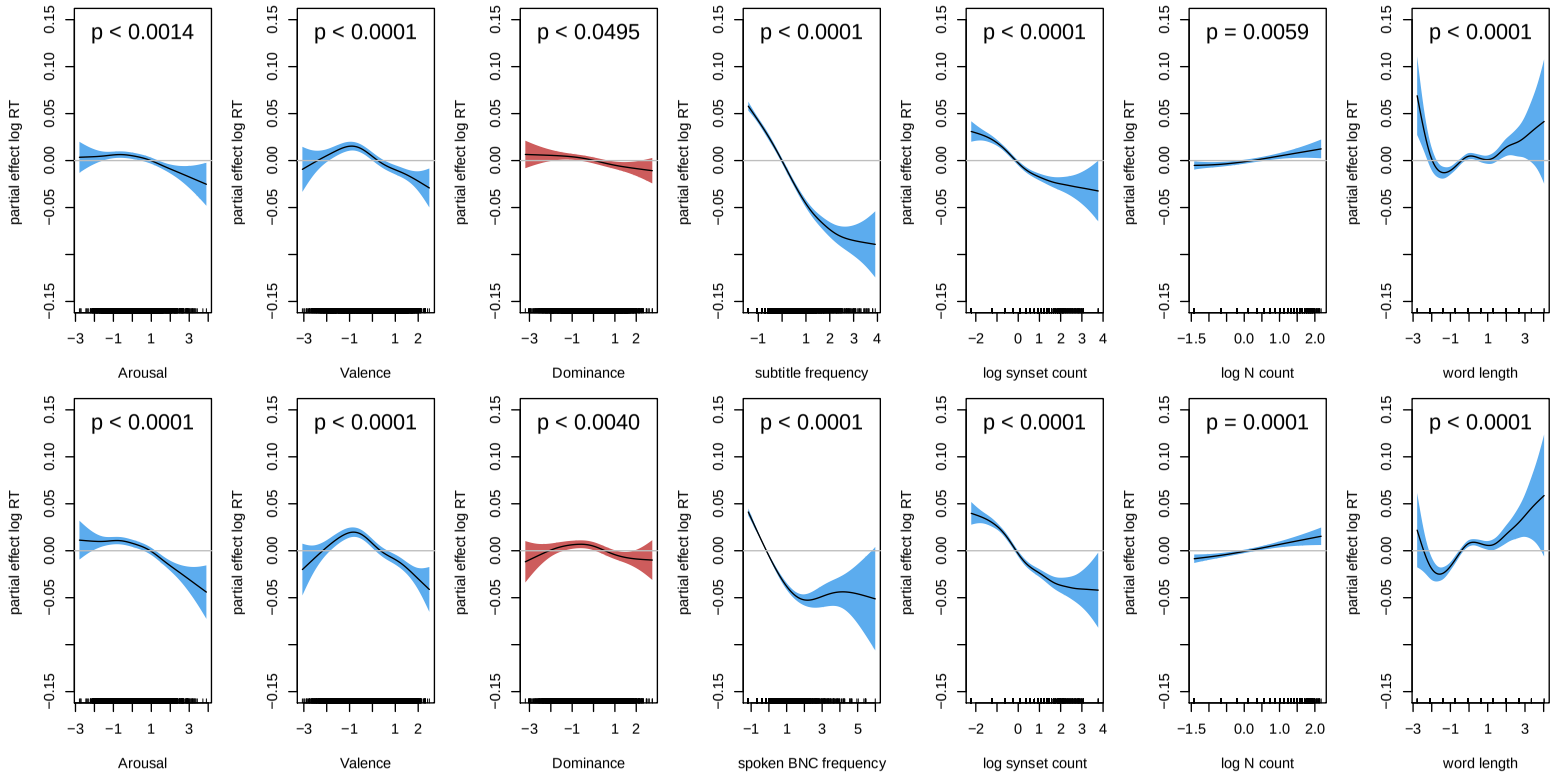


Figure 3: Partial effects on log RT for number of synsets, number of orthographic neighbors, word length (in letters), frequency, arousal, valence, and dominance. The top panels represent a model using subtitle frequency, the bottom panels represent the corresponding model with BNC spoken frequency.

words are presented and processed in a different register (reading words as they appear in a textual, fictional discourse), the superiority of subtitle frequencies for modeling lexical data disappears as predicted.

# 3 Frequency and individual experience

## 3.1 "Average samples" and individual experience

Corpora are samples of — usually a variety of — registers of speech or text produced in a language community, and representing a sample of the usage common in that community. This kind of compiled corpus is not, however, a good model for the experience of individual speakers, because language usage is more varied across individuals than corpora tend to imply. For example, research on authorship attribution has uncovered that writers, and even non-professional writers, have their own characteristic habits of word use, which is tuned differently across registers (Baayen, Van Halteren and Tweedie, 1996; Halteren et al., 2005).

To begin to understand why individual language experiences vary so much, it is worth realizing that the number of words any individual can sample over a lifetime is highly restricted. Someone encountering 2 words per second night and day for 80 years would experience around 5 billion word tokens across her lifespan. We might consider a figure in this ballpark to represent the upper bound of possible human linguistic experience. A more realistic estimate, although in all likelihood still far too high, would be to reduce this number by a third, assuming eight hours of sleep. If we then consider a twenty-year old participant in a psycholinguistic experiment, and assume a rate of experience akin to this second guesstimate, the number of words we might expect them to have experienced would be around 840 million. This represents a cumulative experience that is roughly twice the size of the COCA corpus, and less than our 1,100 million subtitle corpus. Accordingly, it seems clear that many of the corpus resources currently available sample more linguistic experience than any individual will, and that any individuals linguistic experience is correspondingly far sparser.

Moreover, it is likely that no twenty-year old, and in fact, probably no other individual native speaker of English, has the exposure to the sheer variety of texts that are sampled in carefully curated corpora such as Brown, BNC and COCA. As a word's frequency decreases, it becomes more likely that exposure to this word is limited and ever more specific to a particular domain of experience and a smaller group of speakers. What this means is that while higher-frequency words are known by all speakers, as we move down to the lower-frequency words, usage fractionates across the population. Gardner et al. (1987) illustrated this phenomenon by testing a group of nurses and a group of engineers on common and occupation-specific vocabulary. As expected, nurses responded more slowly to terms specific to engineering, and the engineers had trouble with words specific to the health care sector. (A large crowd-sourcing lexical decision experiment by Keuleers, Stevens, Mandera and Brysbaert (2015) serves to underline the importance of the relationship between the prevalence of lexical knowledge and lexical processing: as the proportion of speakers who correctly distinguish words from nonword foils decreases, reaction times and error rates increase.)

A further factor that shapes individual linguistic experience is a well-known property of word occurrences known as *burstiness* (Church and Gale, 1995): Once a topic is broached, words pertaining to that topic will be used and re-used with greater than chance probability. Taken together with the factors noted above, this in turn means that while high frequency words will be experienced at a rate that is roughly equivalent to their average rate in a corpus across time and/or individuals, as word frequencies decrease, the chance of a given word being encountered at a given time or by a given individual will drop far below the rate suggested by its average frequency in a large corpus, and in situations where that word actually is encountered, it will tend to be experienced by indi-

viduals at a rate far above that suggested by its average corpus frequency. A consequence of this is that speakers who know a particular low-frequency word will use that word more often than the frequency count itself suggests. And this compensates for their non-use of vocabulary, unknown to them but present in the corpus, that is particular to other individuals' experience and expertise.

A straightforward consequence of the burstiness of word use and of speakers' experiental specialization, and the concomitant fractionation of vocabulary knowledge within society, is that when corpora sample texts covering many registers and many topic domains, words will show a non-uniform distribution across these texts. Following work in statistics (Johnson and Kotz, 1977), the number of different texts in which a word occurs is known as its *dispersion* (Baayen, 1996; Gries, 2008, 2010). A word that consistently occurs across many texts is not only more likely to be a basic word (Zhang, Huang and Yu, 2004), but will also tend to be a word with multiple meanings and many different senses.

In psychology, dispersion is also known as contextual diversity, and it has been argued that once contextual diversity is taken into account, word frequency as such is no longer predictive in tasks such as visual lexical decision and word naming (Adelman, Brown and Quesada, 2006). However, it is interesting that Heister and Kliegl (2012) report that dispersion failed to have predictive power for German data, and Pham (2014) reports similar results for Vietnamese. However, as we hope the foregoing has made clear, not only does the notion of "lexical frequency" raise questions about *what to count?*; where, exactly, counts are drawn from, and what, exactly, they are intended for are also critical areas of concern.

To try to establish which of these factors might account for the different effects of contextual diversity observed in English on one hand, and German and Vietnamese on the other, we return to the reaction time data for the set of 4440 English words we examined earlier. To initally see whether dispersion did indeed provide a better account for these data, we compared two sets of frequency and dispersion measures, one pair drawn from the Ghent subtitle corpus, and the other pair drawn from the BNC. In both models, the dispersion measures failed to reach significance ($p > 0.1$); by contrast, the frequency measures revealed the usual huge effect sizes.

In other words, in neither corpus, each of which is standardly used as a source of psycholinguistic metrics, did we find that contextual diversity was a better predictor of behavior than lexical frequency. Why? The obvious answer is, as we noted above, that *where* one takes counts from is as important as *what* one counts.

Adelman et al. (2006) based their initial analysis on a subset of the TASA corpus, which contains short excerpts from texts appearing on the curriculum of high-school students, reflecting the different subjects in which this population is educated. For all of the reasons we have described above, the distributional properties of the words in this very specific set of corpus materials can be expected to differ in a variety of ways from subtitle English, normal conversational English, and standard written English as sampled, e.g., by the British National Corpus, especially with respect to the balance between frequency of occurrence and measures of semantic richness. And it seems clear that whether or not support for dispersion as a predictor of lexical processing is or is not found in an analysis can ultimately depend on which of these corpus resources one selects.

### 3.2 Sampling the experience of the individual

A final, fundamental, problem we should highlight in relation to the question of measuring the effects of lexical frequency is that of accounting for the way that the the statistical properties of language serve to influence the experience of individual speakers over their lifetime is subject to continuous change.

In the earliest years of life, children learn from their parents and their peers, and from the very

outset, individual circumstances contribute in turn to the amount and the variety of linguistic input that individual children experience: differences in the amounts parents talk (Hurtado, Marchman and Fernald, 2008; Weisleder and Fernald, 2013), in socioeconomic status (Fernald, Marchman and Weisleder, 2013), and even the quality of their day care (Stolarova, Brielmann, Wolf, Rinker and Baayen, 2015) all result in measurable differences in lexical development. As children then progress through the educational system, their experiences of the language will further diversify along with the more specific education received. And when they become parents themselves, words that were frequent in early childhood (*nappy, bib*) and that fell into disuse, may once again come back into frequent use, a cycle of use and disuse that may repeat itself when they become grandparents.

In addition to social and biologically-driven cycles in words' frequencies, further shifts in lexical knowledge may result from changing occupations, traveling or moving to other places, meeting new people or simply watching TV. Indeed, the distribution of lexical items essentially guarantees that throughout their lifespan, any speaker that continues to engage with language will continue to learn new words (Ramscar, Hendrix, Love and Baayen, 2013; Keuleers et al., 2015). The same holds for their knowledge of the patterns of lexical co-occurrences (and non-occurrences that characterizes any linguistic system as a whole, Ramscar et al., 2014).

Figure 4 illustrates the latter finding. The response variable is accuracy in paired associate learning (PAL). Discrimination learning theory (Ramscar, Yarlett, Dye, Denny and Thorpe, 2010; Ramscar, Hendrix, Love and Baayen, 2013) predicts that learning to associate a pair of words will depend on at least two simple counts. First, the more often the words co-occur together, the better subjects should be able to recall the second word given the first. Second, the more often the first word occurs without the second word, the worse performance should be. Figure 4 shows that these predictions are supported by the data: beta weights are positive for cooccurrence frequency, and negative for the frequency difference. Of interest here is how these coefficients change over the lifetime: As experience accumulates over the lifetime, the (absolute) magnitude of the coefficients increases. This indicates that the older a subject is, the more this subject is "sensitive" to these frequency measures. Another way of expressing this finding is that the older a subject is, the more they are attuned to the *systematic* effects of the patterns of lexical co-occurrences in the system of speech in their community — as we grow older, we have had more opportunity to sample word use in our speech communities, and this is reflected in a deeper knowledge of the systematic properties of a language.

The accumulation of knowledge over the lifetime comes with a cost. As vocabulary knowledge increases in adulthood, the entropy (the average amount of information) associated with this knowledge will also increase, causing processing speed to decrease (Ramscar, Hendrix, Love and Baayen, 2013). The balance of knowledge and speed is beautifully illustrated by lexical decision RT and accuracy: Older subjects respond more slowly, but with much greater accuracy. Indeed, for the lowest frequency words in the data set studied by Ramscar, Hendrix, Love and Baayen (2013), young subjects' responses are almost at chance, whereas even for the hardest words, older respondents are 80% correct.

The inevitable increases in lexical entropy brought about by continuous sampling across the lifetime are further reflected in other changes in linguistic behavior. For example, the use of pronouns instead of personal names increases as adults age (Hendriks, Englert, Wubs and Hoeks, 2008), and this can be seen as compensatory strategy to help deal with the processing demands inevitably imposed by the entropy of personal names, which increases dramatically across the lifespan (Ramscar et al., 2014). Interestingly, this change is not just an adaptation of the individual. The number of personal names in use in English has itself increased exponentially since the Victorian era (Ramscar, Smith et al., 2013), and the same pattern of increase in the use of personal pronouns in lieu of personal names has also been observed in the English language itself, as it has developed over the

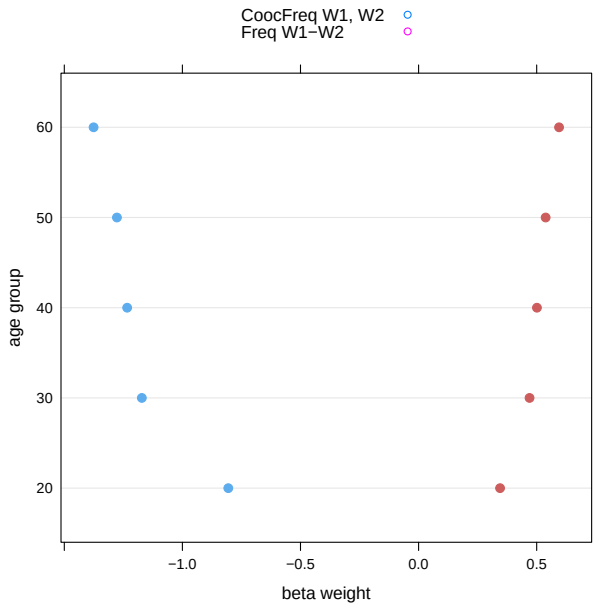


Figure 4: Coefficients (beta weights) in a model predicting accuracy in paired associate learning from the cooccurrence frequency of the two words, and the extent to which the frequency of the first word exceeds that of the second word.

last 200 years (Baayen, Tomaschek, Gahl and Ramscar, 2015).

## 4 Objective versus subjective frequencies: the "age of acquisition"

Our analysis of the PAL task (plotted in Figure 4) shows not only that language learning continues throughout the lifespan, but also that this process is highly systematic: lexical items are not learned in isolation, but rather, learning of any given item can impact any other item that it has been explicitly or implicitly related to by experience. This in turn raises a question: To what extent does it matter how *long* a word has been in a speaker's vocabulary? The age at which a word is learned has often been put forward as an important determinant of lexical processing (Carroll and White, 1973a, 1973b). Indeed, although some studies suggest that frequency of occurrence and age of acquisition are both important predictors in their own right (see, e.g. Brysbaert, 1996; Brysbaert, Lange & Van Wijnendaele, 2000), it has been argued that once age of acquisition is taken into account, frequency of occurrence is no longer a significant predictor (see, e.g. Morrison & Ellis, 1995; Barry, Hirsh, Johnston & Williams, 2001).

Regardless of the merits of the specific nature of these claims, age of acquisition has asserted itself as a much stronger predictor than word frequency across many tasks (Ghyselinck, Lewis and Brysbaert, 2004; Cortese and Khanna, 2007). To explain the strong effect exerted by age of acquisition in chronometric tasks, it has been proposed that these effects are the consequence of loss of neuronal plasticity over the lifetime (Ellis and Lambon Ralph, 2000).

However, there are some good reasons why the idea that the age at which a word is acquired should prove to be a strong determinant of adult lexical processing ought to be viewed sceptically.

First, recent studies (see, e.g., Hofstetter, Tavor, Moryosef & Assaf, 2013) indicate neuronal plasticity is much more pervasive than previously assumed. Furthermore, as outlined above, positive evidence is now available that lexical knowledge accumulates with the years, leading to greater

vocabulary size as evidenced by more accurate lexical decision performance (Keuleers et al., 2015), and enhanced sensitivity to the distributional properties of the language (Ramscar, Hendrix, Love and Baayen, 2013; Ramscar et al., 2014).

Second, while discrimination learning models have proven to be successful at both fitting and predicting changes in the systematic relationships between lexical items across the lifespan and local sequence effects in learning (Jones, Curran, Mozer and Wilder, 2013), as well as a large range of other phenomena in human and animal learning (see Ramscar, Dye and McCauley, 2013 for a review), these results are themselves incompatible with the existence of very large and pervasive effects of learning order that apply over and above more usual systemic learning factors, as age of acquisition effects would appear to indicate.

Third, it is not straightforward to obtain good measures of age of acquisition. Proponents of age of acquisition as a causal factor shaping lexical memories take the considerable convergence of child-elicited data and adult judgements on one hand, and the predictivity of these age of acquisition measures for chronometric measures of lexical processing on the other hand, as evidence of the validity of the age of acquisition construct. Unfortunately, although one can ask children to name pictures and derive an 'objective' measure of age of acquisition from their responses (Gilhooly and Gilhooly, 1980; Goodman, Dale and Li, 2008), this does not do justice to children's ability to understand words long before they are able to articulate these words themselves. Alternatively, one can ask educators about the age at which they believe children know words (Brysbaert, 1996; Morrison, Chappell and Ellis, 1997), or one can consult non-specialists by means of crowd-sourcing to evaluate the age at which they believe they would understand a word if it was uttered (Kuperman, Stadthagen-Gonzalez and Brysbaert, 2012). However, it is doubtful that adult judgements can reflect the true age at which words are 'first acquired', for two reasons. First, age of acquisition ratings from adults may be influenced by the ease of information uptake from the visual input. Second, it can be ruled out that adults have access to childhood memories of when they first understood a given word. As a consequence, age of acquisition ratings are measures based on a mixture of high-level common-sense reasoning about what words children might know and low-level intuitions.

The plausibility and veracity of judgements of age of acquisition are challenged both by the hugely unreliable nature of childhood memories, and by the pattern of development of the neural circuits that ultimately underpin later subjective judgements of age of acquisition when they are elicited.

The idea that adults have virtual amnesia for their childhood experiences is an old one (Henri and Henri, 1895; Freud, 1905/1953), and studies have shown that later memories for events in the first decade of life are both vague and intermitent (Bauer and Larkina, 2014), with recall for events prior to age 3 being largely non-existent: a meta-analysis of a range of methods aimed at probing childhood memories found the mean age of earliest memories of *anything* dated to 3.4 years of age. Moreover, these early childhood memories, which are best characterized as "fragmentary, disorganized, and often enigmatic in the sense that the rememberer does not know why (or how) they remember them" (Wells, Morrison and Conway, 2014), are also extremely unreliable: many will later be rejected as false, either because the event in question occurred to someone else, because they are corrected by a parent or peer, because later experience reveals them to be implausible, or because they conflict with later established facts (Mazzoni, Scoboria and Harvey, 2010).

In the first 4 years of life, childrens long term autobiographical memory capacities appear to be negligible (Scarf, Gross, Colombo and Hayne, 2013). In later childhood, when very young children are asked to retrieve autobiographical memories, their behavior differs substantially from that of adults Wells et al., 2014. It appears that childrens autobiographical memory systems only begin to approximate those of adults some time between 10 to 15 years of age (Van Abbema and Bauer,

2005; Conway, 2005), and that full adult memory capacity does not emerge until late adolescence and early adulthood (Habermas and Bluck, 2000; Ghetti and Bunge, 2012).

This pattern is striking when one considers the development of the neural circuits that appear to regulate the encoding and retrieval of autobiographical memories. A range of neuroscience measures reveal that the kind of directed memory retrieval participants are presumed to engage in while recalling the age at which they learned words reliably engages a Ventral Fronto-Temporal Pathway, comprising systems in Prefrontal Cortex (PFC) that support cognitive control, along with systems in the the medial temporal lobes (MTL) and basal ganglia (and especially the striatum) that support the learning and encoding of information in memory (Fletcher, Shallice and Dolan, 1998; Fletcher, Shallice, CD, SJ and Dolan, 1998; Jacques, Kragel and Rubin, 2011; Bekinschtein and Weisstaub, 2014). Notably, although the PFC has been shown to play a critical role in directed memory retrieval (Barredo, Öztekin and Badre, 2013; Peters, David, Marcus and Smith, 2013, the encoding systems of the Ventral Fronto-Temporal Retrieval Pathway and the PFC develop differently. Whereas the MTL and basal ganglia appear to develop early, and to exhibit relatively high postnatal functionality (Chugani, 1996; Bekinschtein and Weisstaub, 2014), there is little to no evidence that much, if any, of the functionality associated with the adult PFC is available to children prior to their fourth year (Ramscar and Gitcho, 2007); and as adult functionality does develop, its timecourse closely mirrors the pattern of behavioral change we described in the development of autobiographical memory (Thompson-Schill, Ramscar and Chrysikou, 2009; Somerville and Casey, 2010; Ghetti and Bunge, 2012).

It thus appears that childrens initial experiences are learned and encoded by systems in which the components that ultimately direct the retrieval of autobiographical memories are undeveloped and unintegrated as compared to the encodings' components, and that retrieval capacities develop and become more integrated relatively slowly throughout childhood and adolescence (Luciana and Nelson, 1998). This developmental perspective on the slow, incremental coupling between memory formation and memory retrieval helps make sense of why, despite the fact that children clearly learn and form memories all the time, it is the case that after a surprise visit to a fire station, 5 to 6 year-olds are later unable to freely recall details such as weather, time of day, duration, etc, (and why this situation barely improves with explicit questioning); and why even 9 to 10 year-olds performance on this task is so surprisingly poor (Strange and Hayne, 2013).

Given the undeniable behavioral and developmental facts about childhood autobiographical memory, we can rule out that adult ratings are influenced by memories of initial lexical acquisition. Furthermore, although it is logically possible that word memories are stronger the earlier they are acquired, this logical possibility is at odds with the development of autobiographical memory. Adding to this that even developmental psychologists are undecided on exactly what it means to 'learn a word' (Tomasello, 2009), it would be truly remarkable if the magic months at which one learns words like *doze* (48 months) versus *wiggle* (96 months, estimates from Kuperman et al., 2012) indeed affect adult reading times.

In what follows, we investigate this issue further by considering the predictive power of age of acquisition ratings for lexical processing measures, using the set of ratings collected by Kuperman et al. (2012) for 30,000 English words. Following pre-established methods (Stadthagen-Gonzalez and Davis, 2006), Kuperman et al. (2012) asked subjects to estimate the age (in years) at which they felt they had learned a given word, and would have understood it if it was uttered. As is well known, ratings of age of acquisition correlate well with other measures of age of acquisition obtained from educators or children (Gilhooly and Gilhooly, 1980; Brysbaert, 1996; Morrison et al., 1997; Goodman et al., 2008), and this set did not differ in that regard.

The point of the following analyses is not to contest these correspondences, but rather to address three questions. First, we are interested in the lexical-distributional properties of words as a function

of when they came into use in the early years of life. Here, we expect very similar results when the present ratings are exchanged for other estimates of age of acquisition that do not depend on ratings or adult expert judgements. Second, since ratings are a human response variable just as lexical decision latencies, we are interested in the 'human factor' in these ratings. That is, to what extent is the production of these ratings determined by the same determinants of lexical processing that they are then used to explain? To what extent are these ratings influenced by the same factors that influence visual uptake in reading the presented words? Are these ratings influenced by semantic similarity to other words in the lexicon, and specifically, to childhood words? And are there dimensions of lexical variation on which ratings and reaction times diverge? Third, to what extent is age of acquisition a causal factor underlying frequency of use, and to what extent is it a caused factor that is itself determined by, e.g., frequency of use?

We address this third question below in the section on graphical modeling. To address the first two questions, we begin with considering what is involved when asked when to first have understood the word *diaper* (a piece of cloth or other fabric that serves as an undergarment for babies). Given our review of the development of autobiographical memory, we shall assume that the possibility can be ruled out that responses are based on actual recalling a memory of a child first understanding the meaning of *diaper* (at around 41 months, according to Kuperman et al's data). One way in which an adult in the experiment of might approach this task is to reflect about the age at which one sees little kids wearing diapers. A second way in which a rating may be provided is by intuitive guessing. When it comes to words such as *abstract*, *wreckage*, and *windpipe*, pinpointing a specific age of acquisition becomes difficult. This may not only give rise to frustration on the part of the rater, but this frustration may also give rise to inflated ratings for earlier ages of acquisition.

Given the difficulty of the age of acquisition rating task, it is not surprising that the rated age of acquisitions often miss their target. Empirically, an examination of the CHILDES corpus establishes that all of the words analyzed in the present study are commonly used by caregivers addressing children between 2 and 6 years of age. However, the mean age of acquisition for these words in Kuperman et al's ratings data is 7 years, with a range from 2.2 to 14.8.

To assess the role of intuition and possibly frustration when rating for age of acquisition, we calculated the cosine of the angle between each of the words in our data set and 12 'pivotal' words that we expected to influence rating behavior. First, we included words for the young (*baby, boy, girl, young, child*) and their caregivers (*mum, daddy*). In addition, we included *play* and *toys* as childhood words. Since in English *baby* is also used to address adults, we included the words *sex* and *sexy* to create a reference level that is highly unlikely to be a dominant concept in early language acquisition. Finally, to gauge participant frustration, we included the expletive *fuck*.

Our hypothesis is that in comprehension, meanings of words other than the target word co-resonate (cf. Bowers, Davis & Hanley, 2005) to different degrees depending on the extent to which the visual input is discriminative on the one hand, and on the distributions of collocations on the other (for detailed discussion, see Shaoul, Willits, Ramscar, Milin & Baayen, 2015). Thus, when being confronted with the written word *windpipe*, we expect raters to be influenced by the degrees to which the semantics of *windpipe* is similar to the meanings of our pivot words.

Semantic vectors were obtained from the British National Corpus using naive discrimination learning (see Shaoul, Willits et al., 2015). Due to limitations on matrix size, our data set was reduced to the 3503 words for which cosine similarities between target words and pivotal words could be calculated. Substantial collinearity ($\kappa = 47$) of the resulting twelve 3503-element vectors of cosine similarities suggested orthogonalization by means of principal components analysis. In what follows, we made use of the first five principal components, as summarized in Table 2. For the interpretation of the components, we inspected which words had extreme loadings on these components. Furthermore, we calculated all pairwise cosine similarities for the pivotal words, which made

| PC | var. prop. | negative loadings | positive loadings |
|---|---|---|---|
| PC1 | 0.44 | | **young, boy, child, girl, baby** |
| PC2 | 0.25 | **mum, daddy** | **toys, young, girl, boy** |
| PC3 | 0.09 | **play** | |
| PC4 | 0.08 | | **fuck** |
| PC5 | 0.05 | **sex**, sexy | daddy, **mum** |

Table 2: Principal components, the corresponding proportion of variance, and words with strong negative or strong positive loadings. For words in bold, the loadings on the pertinent principal component are correlated ($p < 0.05$) with the vector of corresponding cosine similarities.

it possible to correlate the loadings on a given principal component with the vector of corresponding cosine similarities. Words for which a significant ($p < 0.05$) correlation was observed are printed in bold in Table 2.

The first dimension represents words for the young (*baby, girl, child, boy, young*). The second principal component contrasts the parents (*mum, daddy*) with the young (*boy, girl, young, toys*). The third dimension singles out *play*. The expletive *fuck* determines the fourth dimension. To verify that *fuck* is indeed capturing frustration, we also calculated its cosine similarity to *damn* and *shit*. These cosine similarities were larger than for any of the other 11 pivotal words, including *sex* and *sexy*. The final principal component contrasts *mum* with *sex*.

Given the foregoing, we expected that these five new predictiors would explain more additional variance for the ratings than for the reaction times, as the pivotal words are chosen specifically to target the focus on 'early' semantics. We also expected that they would explain the variance in the RTs, even in the presence of Age of Acquisition as predictor. Finally, it is also likely that differences between individual PCs may emerge across tasks: For instance, words with semantic vectors similar to those of words expressing sexuality may elicit shorter reaction times from adult subjects, but perceptive participants rating for age of acquisition may decide that sexuality is irrelevant for early word learning, and opt to rate them as having a later age of acquisition.

To test our predictions, we fitted generalized additive models to both the age of acquisition ratings and the reaction times, including the five pivotal principal components as additional predictors. As expected, adding the PCs to the model for the age of acquisition ratings increased the adjusted R-square by 0.045 from 0.297 to 0.342 (decrease in fREML (fast RElativized Maximum Likelihood) score, a measure of goodness of fit, 99.98, for 10 degrees of freedom, $p < 0.0001$, smaller fREML scores indicate improved fits). For the reaction time, the benefit of adding the five PCs was a smaller increase of 0.013 from 0.403 to 0.416 (decrease in fREML 21.8 for 9 degrees of freedom, $p < 0.0001$).

Figure 5 presents the partial effects of the predictors for Age of Acquisition, and Figure 6 visualizes the partial effects for the model for the reaction times. Model summaries can be found in the appendix. For both models, a nonlinear interaction was present involving frequency and PC1. These interactions, modeled with tensor product smooths[5], are presented in Figure 7.

With respect to rated age of acquisition, we find, as expected, that more frequent words are judged to be learned earlier, that longer words are rated as acquired later, and that words with positive valence and many meanings and senses are judged to be early words.

A further surprising effect is that of neighborhood density (N Count): when people are asked to introspect age of acquisition, it would seem that greater densities lead to earlier estimations. At first

[5] A tensor product smooth uses higher-dimensional simple basis functions (restricted cubic splines) to set up a mesh for a wiggly surface. Again, penalties on wiggliness guarantee an optimal balance between over- and undersmoothing.

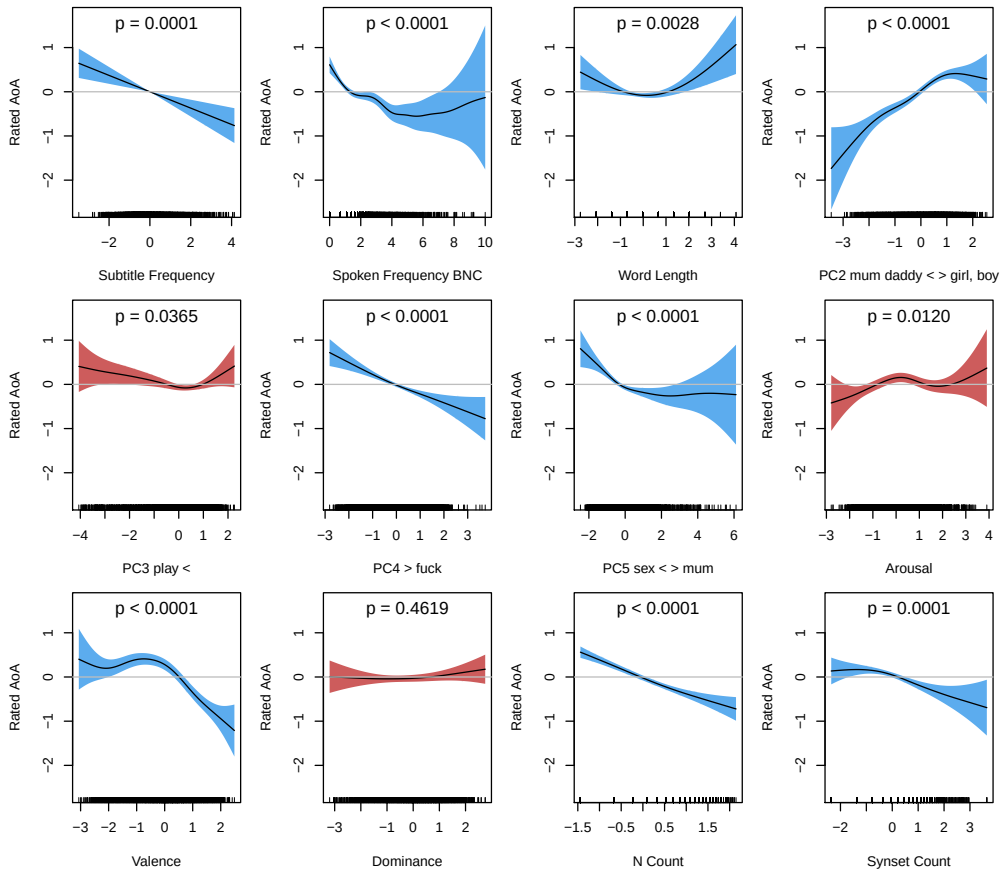


Figure 5: Partial effects of predictors for rated Age of Acquisition. Predictors without good support ($p > 0.01$) are shown in red. For the effect of PC1, see Figure 7.

sight, this might seem to contradict other findings, since in both reading and speech production, a greater neighborhood density has been argued to be an indicator of delayed processing (Balota, Cortese, Sergent-Marshall, Spieler and Yap, 2004; Sadat, Martin, Costa, Alario et al., 2014).

However, it is *also* the case that high neighborhood density correlates well with the presence of high-frequency digraphs that are shared with many other words (Nusbaum, 1985; Baayen, 2001). Furthermore, orthographic neighborhood density likely also correlates with the presence of high-frequency diphones. Importantly, words that share many diphones (or letter pairs) are more likely to be (i) higher-frequency words and (ii) words that have more similar articulatory gestures than words that share few or no diphones. Thus, what we are finding here likely reveals that early words tend to be words with simple and well-practiced articulatory gestures. Accordingly, our interpretation of this finding is that this is a genuine effect, and that it reflects a simple communicative constraint: Unlike words learned in later life, the vocabulary people employ when talking to young children is biased towards words that are (and should be) relatively easy to articulate by the young.

All five of the pivotal PCs appear in the model. PC1 is discussed below. Somewhat to our surprise, the support for PC3, although nominally just significant, is too weak to support further interpretation — apparently, playing is less characteristic for the early years than we had thought, perhaps because adults engage in playing both with children and amongst themselves (both in sports and in board games). The effect of PC2 suggests that words that have semantic vectors similar to parents as caregivers are judged to be early words. The effect of PC4 may reflect a frustration that

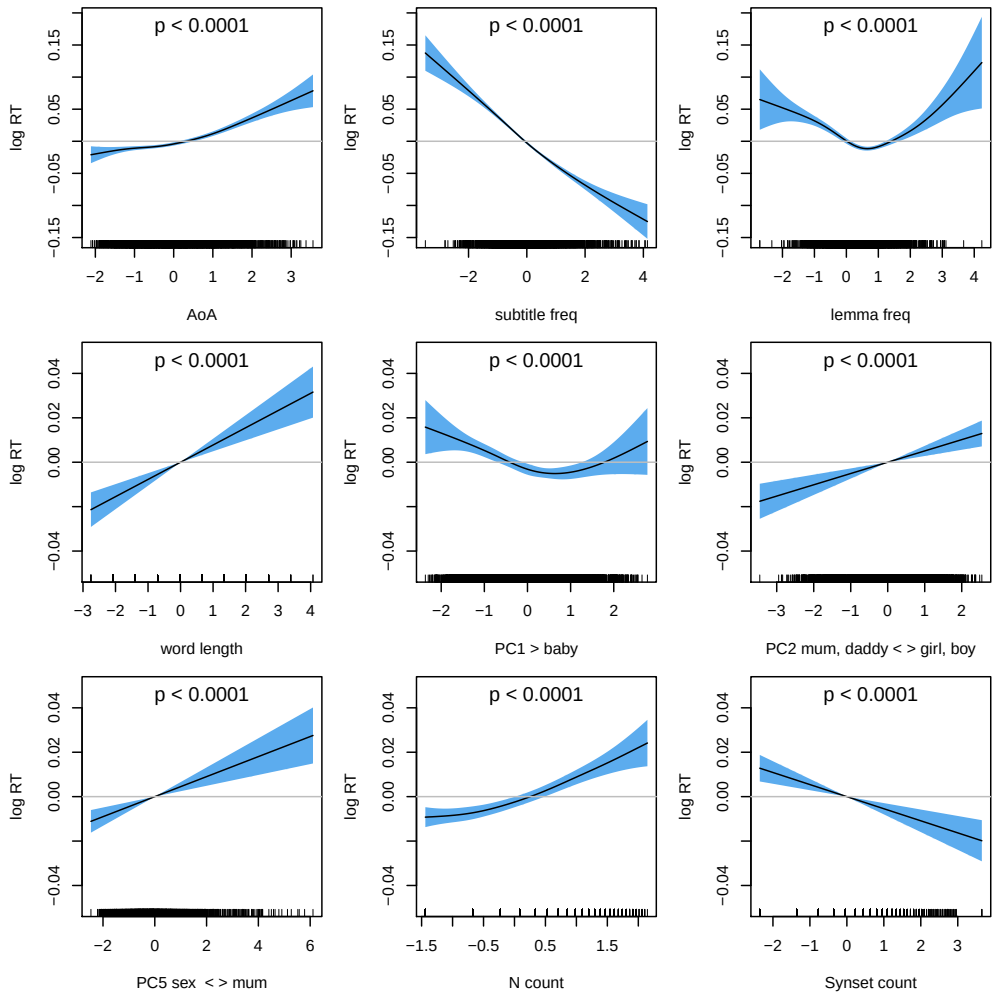


Figure 6: Partial effects of predictors for log RT. Upper panels have the same scale, for the remaining panels with smaller effect sizes, a larger scale is used. Predictors without good support ($p > 0.01$ are shown in red. The effect of PC1 is its partial effect, for its effect in interaction with subtitle frequency, see Figure 7.

raters are likely to feel during the task of rating for age of aquisition, since it shows that the more similar a word is to the expletive *fuck*, the earlier that word is rated. Accordingly, it seems that cuing a typical adult outlet for frustration leads raters to spend less time on this task, giving rise to the production of an earlier rating than in other circumstances. Finally, PC5 indicates that, as expected, words semantically close to sex, independently and opposite to the mother as prototypical caregiver, are rated as late words.

Next we consider the reaction times themselves. As expected, a lower rated age of acquisition predicts shorter RTs. Compared to the effect of subtitle frequency, the effect of rated Age of Acquisition is modest. Lemma Frequency shows a U-shaped effect, which may indicate that participants expect words to be of average frequency, and pay when this expectation is not met. Longer words elicit longer RTs, as do words with more orthographic neighbors. More meanings and senses afford shorter latencies.

Three of the pivotal PCs are well-supported as predictors of response latencies. Words with semantic vectors closer to the semantic vectors of words for the young give rise to shorter RTs. The

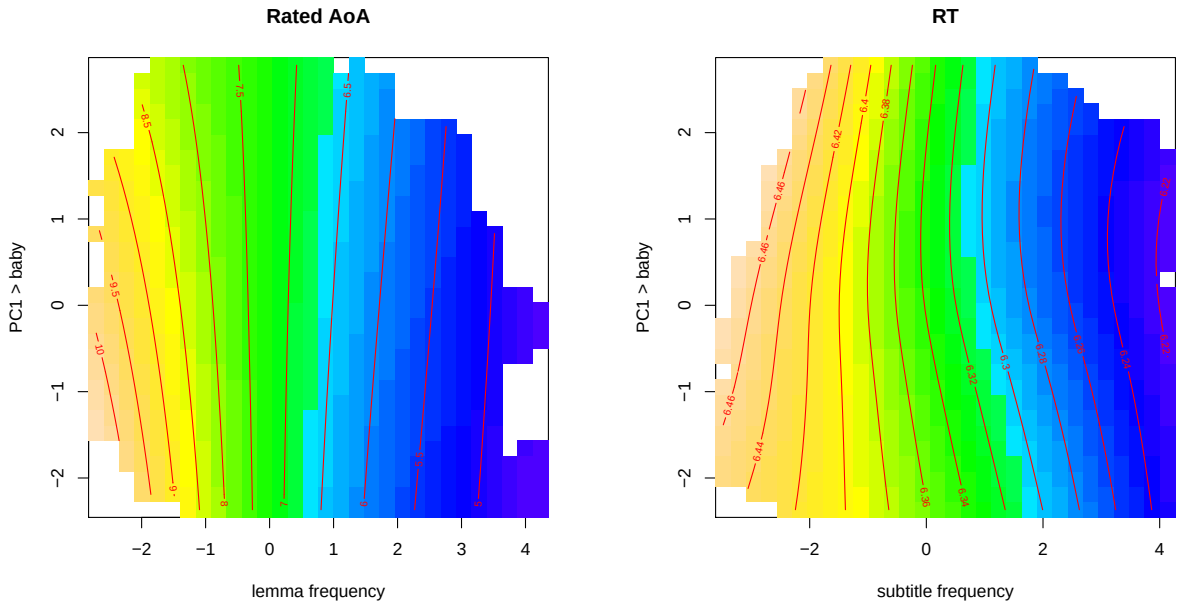


Figure 7: Tensor product smooths for the interaction of lemma frequency and PC1 (left) in the model for rated age of acquisition, and for the interaction of subtitle frequency and PC1 (right) in the model for log reaction time. Larger values of PC1 correspond to greater cosine similarity with *baby, child, girl, boy*, and *young*. Orange colors indicate large values, blue colors low values of the response. (The prediction surfaces are adjusted for the most typical values of the other predictors in the model.)

effect graphed here is the partial effect of PC1, its interaction with subtitle frequency is discussed below. PC2, highlighting *mum* and *daddy* as caregivers of the young, affords shorter RTs. The effect of PC5 is opposite to that observed for age of acquisition, with semantic proximity to *sex* giving rise to shorter response times. We may be dealing here with facilitation from a subtle form of sexual arousal. (Note that Arousal is not predictive given the other variables in the model.) For PC5, however, age of acquisition raters are not led astray, and properly adjucate sexual words to later years.

Thus, PC5 and neighborhood density provide us with a window into where it is that rated Age of Acquisition and reaction time in the lexical decision task fundamentally diverge.

Figure 7 visualizes an interaction of frequency and PC1 that was present for both response variables.[6] A greater lemma frequency (ratings) and a greater subtitle frequency (RTs) goes hand in hand with early acquisition and short response times.

For the ratings, we find that words dissimilar from words for the young (large negative values on the vertical axis) show a large effect of lemma frequency (many contour lines are crossed when moving along the horizontal axis, indicating a steep gradient). Words with a low lemma frequency (in the left-hand side of the plot) show an effect of PC1, with age of acquisition being judged as earlier in life when words are more similar semantically to words for the young, as expected. As lemma frequency increases, this effect disappears.

For the latencies, closer semantic similarity to words for the young predicts longer reaction time

[6] For the reaction times, we decomposed the interaction of frequency and PC1 into two main effects and the remaining joint effect. For the ratings, a similar decomposition resulted in an inferior fit, which is why a non-decompositional tensor product smooth is reported for this response variable.

for low-frequency words. As frequency increases, the effect reverses into facilitation. Approached from the perspective of frequency, we see contour lines that are far apart (low gradient) for low values of PC1, and close together (high gradient) for high values. In other words, words similar to words for the young show a stronger frequency effect. The elongated latencies for low-frequency words with semantics similar to the semantics of words for the young may reflect a processing conflict, with a low frequency of use arguing against a positive lexical decision but semantic proximity to young words arguing in favor of a positive response.

Thus, to summarize, this analysis indicates that age of acquisition ratings and reaction times are co-determined by a wide range of variables, many of which are predictive for both in a consistent way, but other which show divergence. (This is only natural, since lexical response tasks essentially aim to tap 'bottom up' processing, whereas ratings engage 'top down' reasoning in addition to 'bottom up' information uptake.)

## 5 Frequency and learning

The most common term for referencing the basic resource that is assumed to be involved in lexical processing, *the mental lexicon*, bears witness to the pervasive influence that the classical dictionary has as a metaphor for theoretical reflection. Dictionaries list words, enriched with information about their pronunciation, conjectures about their interpretation, and sometimes indications of their frequency of use. Influential models of lexical processing, irrespective of whether they make use of interactive activation (McClelland and Rumelhart, 1981; Coltheart, Rastle, Perry, Langdon and Ziegler, 2001; Dijkstra and Van Heuven, 2002; Taft, 1994) or spreading activation (Levelt, Roelofs and Meyer, 1999) have adapted themselves to this metaphoric structure: They all comprise an inventory of lexical entries, conceptualized in the form of nodes in a network. Typically, these nodes are associated with real numbers proportional to the frequency of occurrence of the orthographic forms that correspond to each entry.

These kinds of models posit an important devision of labor for different lexical properties. On the one hand, lexical access is determined by the network layout, by how nodes are connected and (co)activate each other. Thus, the effect of word length is determined by the number of lower-level nodes linking up to a given word form node, and likewise the effect of number of neighbors arises from the number of strongly co-activated competitor word nodes. On the other hand, constructs such as resting or threshold activation levels are introduced to account for effects of frequency of occurrence. Thus, a crucial design decision in these models is to impose a fundamental separation between the way effects of frequency of occurrence are accounted for, by means of 'counters in the head', and the way the effects of lexical properties such as length and neighborhood density are dealt with, through the network architecture. It is this separation of frequency from other lexical properties that has paved the way for research programs seeking to find the optimal frequency counts for predicting reaction times: The layout of the network is known, the consequences of interactive activation between nodes have been worked out, and what remains to be done is to get the values for the counters in the head just right.

There are two fundamental problems with these kinds of models of lexical processing. First, this class of models doesn't learn. They posit rich hierarchically structured networks, with hand-crafted connection weights, without providing any account of how network structure or connection weights come into existence. Second, the models have no way of accounting for paradigmatic effects in lexical processing.

Paradigmatic effects in lexical processing were first demonstrated by Milin, Filipović Durdević and Moscoso del Prado Martín (2009) for Serbian nominal paradigms, and replicated in subsequent

experiments reported in Baayen, Milin, Filipović Durdević, Hendrix and Marelli (2011). The latter study demonstrated in addition that English nouns are subject to paradigmatic effects at the level of prepositional phrases. Given a particular noun's vector of relative frequencies of use of case endings (Serbian) or prepositions (English), and given the corresponding general relative frequencies of use of these case endings or prepositions, the distance between such specific relative frequencies and the corresponding general frequency distributions can be evaluated with the Kullback-Leibler divergence, also known as relative entropy. This distance measure has been found to be a further co-determinant of response latencies. What these findings show is that how often other case endings and other prepositions, not present in the visual input, are used, predicts response latencies in tasks presenting words in isolation. Such effects are outside the scope of interactive activation models, as well as of parallel distributed processing (Harm and Seidenberg, 1999) or Bayesian (Norris, 2006) models.

An approach that properly predicts such paradigmatic effects, that avoids the complex hierarchical structures of interactive activation models (see, e.g., Taft, 1994, 2004), and that is based on the simplest mathematical formulation of error-driven learning (Rescorla & Wagner, 1972) was proposed by Baayen et al. (2011). In addition to the abovementioned relative entropy effects, the model accounts for a wide range of other effects in the morphological processing literature, including word frequency effects and morphological constituent frequency effects.

At the heart of the model is a simple two-layer network in which input units representing *cues* (discriminative predictors) are connected to output units representing discriminated experiences, henceforth *outcomes*. The weights between cues and outcomes are estimated with the learning of Rescorla and Wagner. When a cue correctly predicts an outcome, the weight on its connection to that outcome is strengthened. The extent to which this weight is strengthened depends on the other cues in the input. The more such cues are present, the smaller the amount is by which a weight is increased. When a cue is present but a given outcome is not, the weight from the cue to this outcome is weakened. When there are more cues in the input, the amount by which the weight is decreased is larger.

As for any computational model, a lot depends on the choice of input and output units. A first option (used by Baayen et al., 2011) is to take letter pairs as cues, and as outcomes those experiences for which we have evidence that they are discriminated by speakers in the form of distinct words. Following Milin, Ramscar, Baayen and Feldman (2015) and Baayen, Shaoul, Willits and Ramscar (2015), we refer to these outcomes as *lexomes*. Lexomes are theoretical constructs, akin to atoms in physics, the activation of which (by the cues in the input) provides excellent predictions for experimental response variables, ranging from reaction times (see, e.g., Pham & Baayen, 2015) to eye-tracking measures to the electrophysiological response of the brain to linguistic stimuli (Hendrix, 2015). Lexomes do not represent word forms, and they also do not represent word meanings. To explain their place in our theory, a metaphor due to de Saussure (De Saussure, 1966) is helpful. De Saussure compared language to a game of chess, in which the value of one particular piece is dependent not only on its own position, but also on the positions and values of all the other pieces that are on the board. A pawn by itself has no independent meaning. A pawn on the second row of a chess board can be totally inert, whereas the same pawn on the seventh row, free to promote to a queen, has tremendous potential. Lexomes are like chess pieces, they contribute to meaning in conjunction with all other lexomes (see Shaoul, Willits et al., 2015 for more detailed discussion).

The *grapheme-to-lexome* (G2L) *activation* is obtained by summing the weights of the orthographic cues in the input (letter bigrams or letter trigrams) to a given lexome. In what follows, the orthographic cues are taken from word triplets rather than from single words The outcomes in this

set-up are the lexomes for the corresponding three words. [7]

For the set of 4440 words investigated above, the correlation of log lexome activation with log subtitle frequency is 0.5: The extent to which a lexome is activated depends on how often it has been encountered, but is at the same time co-determined by the other words that are learned and their orthographic properties. Whereas in the interactive activation frameworks lexical competition is a process that is resolved dynamically through activation and inhibition every time a word is processed, there is no such process in our model: In our approach, competition plays out during learning. Furthermore, effects of frequency of occurrence arise without having to link nodes with counters in the head. The model correctly predicts both whole-word frequency effects and constituent frequency effects for English (Baayen et al., 2011) as well as for Vietnamese (Pham, 2014). Of importance is that in Vietnames, constituent frequency effects are inhibitory, in contrast to the facilitatory effects typically found for English. This difference is captured correctly by the model, which shows that the distributional properties of words in the language, in interaction with fundamental principles of learning, lie at the heart of the observed frequency effects.

A second measure, complementing the input-driven G2L activation, is a lexome's *grapheme-to-lexome prior availability*. This prior availability is estimated with the median absolute deviation (a non-parametric measure of spread) of all the weights on the connections that feed into a given lexome. Lexomes that have more strong connection weights have a higher prior availability. These lexomes are better sustained by the corresponding cues (see Milin et al., 2015, for detailed discussion). For the present data, the prior availability of the lexomes given the orthographic cues (henceforth G2L prior) enters into a stronger correlation with subtitle frequency ($r = 0.63$).

A third measure is derived from a second Rescorla-Wagner network. This network is trained on lexome triplets, predicting the center lexome from the two flanking lexomes. This network provides a second measure of prior availability, henceforth the lexome-to-lexome (L2L) prior, that, instead of being based on the orthography, is grounded in word to word prediction. This second prior also correlates with subtitle frequency to the same extent as the G2L prior ($r = 0.63$) with which it is also correlated ($r = 0.37$). This measure reflects collocational richness, and is similar in spirit to the work by McDonald and Shillcock (2001) and McDonald and Ramscar (2001) on the importance of a word's contextual distinctiveness.

In the L2L learning model, the vectors of weights for a given lexome (as cue) to the lexome outcomes constitute vectors in a high dimensional semantic vector space. Shaoul, Willits et al. (2015) show that semantic vectors based on discriminative lexome-to-lexome learning capture semantic similarities at least as well as other vector space models such as latent semantic analysis (Landauer and Dumais, 1997) and HiDEx (Shaoul and Westbury, 2010). The cosines that we calculated above, in the context of age of acquisition ratings, for our target words and twelve pivotal words (*child, baby, mum, ...*), and from which we calculated the principal components that were used in the analysis of rated age of acquisition, were based on L2L semantic vectors, obtained by training the network on the British National Corpus.

In summary, two networks, one with letter trigrams as cues and lexomes as outcomes, and a second with flanking lexomes as cues for center lexomes, supply us with three further measures, all of which are correlated with frequency of occurrence as well as with reaction times in visual lexical decision (G2L activation: $r = -0.32$; G2L prior: $r = -0.44$; L2L prior: $r = -0.51$).

How important are the measures as predictors of reaction times in the lexical decision task, and for predicting rated age of acquisition, compared to the other measures that we have considered thus far? To address this issue, we made use of a random forest Breiman, 2001; Strobl, Malley and Tutz, 2009; Tagliamonte and Baayen, 2012. Random forests are an excellent choice for assessing the

[7] For details on a model for auditory comprehension, see Baayen, Shaoul et al. (2015).

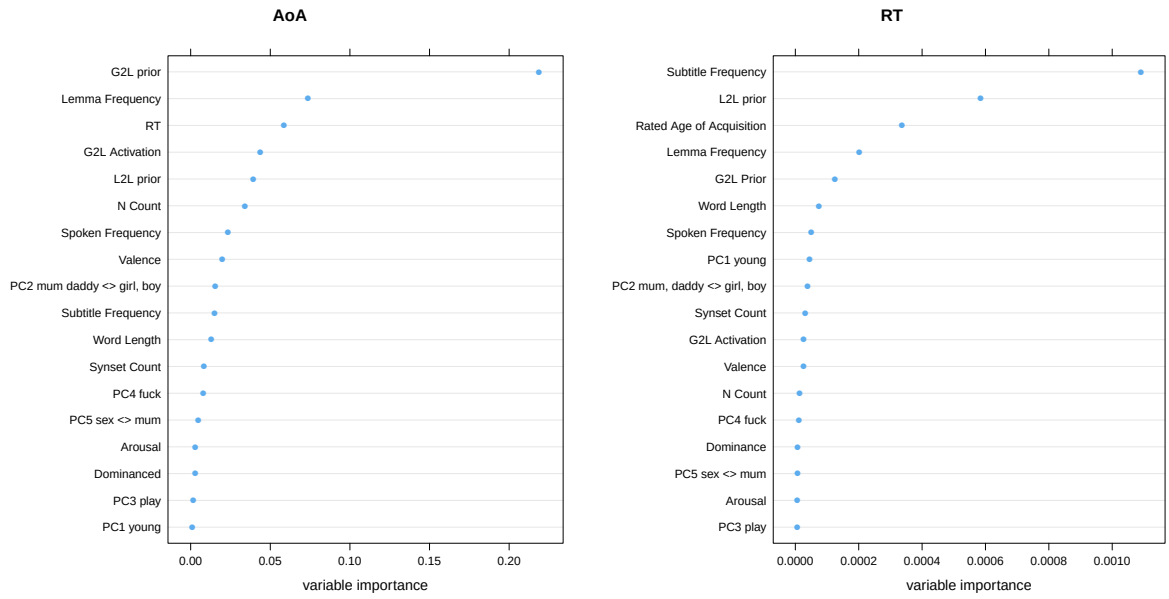


Figure 8: Variable importance of predictors of age of acquisition (left) and log lexical decision response time (right).

relative importance of different predictors, because they are based on a non-parametric recursive partitioning algorithm that is not adversely affected by multicollinearity. We assessed variable importance by randomly permuting the values of a given predictor, thereby breaking the correlation of the predictor with the response. The greater the importance of a predictor, the greater the drop in model accuracy is expected to be when its values are randomly permuted. Figure 8 presents the resulting variable importances, obtained with the `varimp` and `cforest` functions in the `party` package (Hothorn, Buehlmann, Dudoit, Molinaro and Van Der Laan, 2006). For age of acquisition, the most important predictor is the G2L prior, followed at a distance by Lemma Frequency, RT, G2L Activation, the L2L prior, and neighborhood density. Other variables with some importance are Valence, PC2, and Subtitle Frequency. Turning to the reaction times, unsurprisingly, frequency emerges as the most important predictor, followed by the L2L prior and Rated Age of Acquisition. Lemma frequency, the G2L prior, word length and spoken frequency follow, in turn succeeded by the first two principal components of the pivotal words for the young. The remaining variables have smaller and smaller variable importance values, including the classical N-count measure and the three emotion ratings. The squared correlation coefficients for predicted and observed responses for the age of acquisition and reaction time models were 0.64 and 0.67 respectively.

All three learning measures are among the top 5 for age of acquisition (together with lemma frequency and RT), and the two measures of prior availability are among the top 5 in the RT model (together with subtitle frequency, lemma frequency, and rated age of acquisition). It is worth noting that the learning measures are more important for the response quantifying rated onset of learning compared to the lexical decision latencies. The latter may in fact be in part driven by general lexical activation rather than by word-specific lexical discrimination (cf., e.g., Grainger and Jacobs, 1996). The different variable importances shown in Figure 8 suggest that when evaluating age of acquisition measures as a predictor for lexical resilience in aphasia (Brysbaert and Ellis, 2015, this volume), it is worth keeping in mind that it is possible that it may be greater priors or greater valence that underlie this resilience rather than the moment in time that words are supposed to

have been acquired.[8]

## 6 Putting it all together: A graphical model

In the regression and random forests analyses presented in the preceeding sections, we have taken one measure and attempted to predict that particular measure from the other variables available to us. However, correlations between variables are rampant, and to complicate matters further, many variables are the straightforward result of human behavior, not only reaction time or rated age of acquisition, but also spoken frequency, written frequency, and the emotion measures (which are also based on human ratings). In what follows, we therefore change perspective and consider all variables jointly, without selecting one as a "response" and all the others as "predictors", using graphical modeling. A central concept in this approach to finding structure (possibly even causal structure, cf. Pearl, 2009) in a set of correlated measures is *conditional independence*.

Following Højsgaard, Edwards and Lauritzen (2012), let $V$ denote the set of 19 lexical measures, which we now consider as random variables $X_v$ , where $v$ ranges over the different measures (*i.e.*, $\{v \in V\}$). The goal of a graphical model is to simplify the joint density of these variables as much as possible by considering where random variables are conditionally independent. In other words, the joint density is factorized into a product of simpler densities. Let $A$, $B$ and $C$ be disjunct subsets of $V$. Using $f()$ to denote probability mass functions, the random variables $X_A$ and $X_B$ are conditionally independent given $X_C$ ($X_A \perp\!\!\!\perp X_B | X_C$) iff

$$f(x_A, x_B|x_C) = f(x_A|x_C)f(x_B|x_C)$$

(see, e.g., Lauritzen, 1996). For instance, if Age of Acquisition is conditionally independent of Frequency given Valence, N-count, NDL G2L Activation and NDL G2L Prior, this means that if the values of Valence, N-count, G2L Activation and G2L Prior are known, knowledge of Age of Acquisition provides no further information about the value of Frequency.

We made use of the max-min hill-climbing algorithm (Tsamardinos, Aliferis and Statnikov, 2003) as implemented in the `bnlearn` package (Scutari, 2010) for `R` (R Development Core Team, 2015) to estimate the essential graph (also known as the completed partial directed acyclic graph), setting $\alpha$ to 0.000001, and using as test of conditional independence for the associated constraint-based algorithm, the shrinkage estimator for the mutual information (`mi-g-sh`). As results varied with the order of predictors, 30,000 random permutations of predictor ordering were tested, and the model with the minimum deviance (1549.56, for 131 degrees of freedom) was selected. The resulting essential graph, using the `neato` layout of the `Rgraphviz` package (Gentry et al., 2015) is shown in Figure 9.

The `neato` layout brings to the fore the hierarchical structure of the factorization of the joint density

$$f(x_V) = \prod_{v \in V} f(x_v | x_{\mathrm{pa}(v)})$$

for variable sets $\{\mathrm{pa}(v)\}$ ($v \in V$) ($\{\mathrm{pa}(v)\}$ denotes the parents of vertex $v$ in the graph). When two vertices in the graph are not connected by an edge, there exists a set of variables such that given these variables, the two variables associated with the two vertices are conditionally independent. For instance, Subtitle Frequency (`Freq`) and Valence (`Val`) are conditionally independent given Celex Lemma Frequency (`LemF`), Arousal (`Arou`) and the NDL L2L prior (`pL2L`).

[8] Given recent successes of deep learning algorithms in machine learning (see, e.g., Mikolov, Sutskever, Chen, Corrado and Dean, 2013), the fact that a simple three-layer network has problems with learning new words, argued by Brysbaert and Ellis (2015) to support the age of acquisition construct, does not appear to have much force.

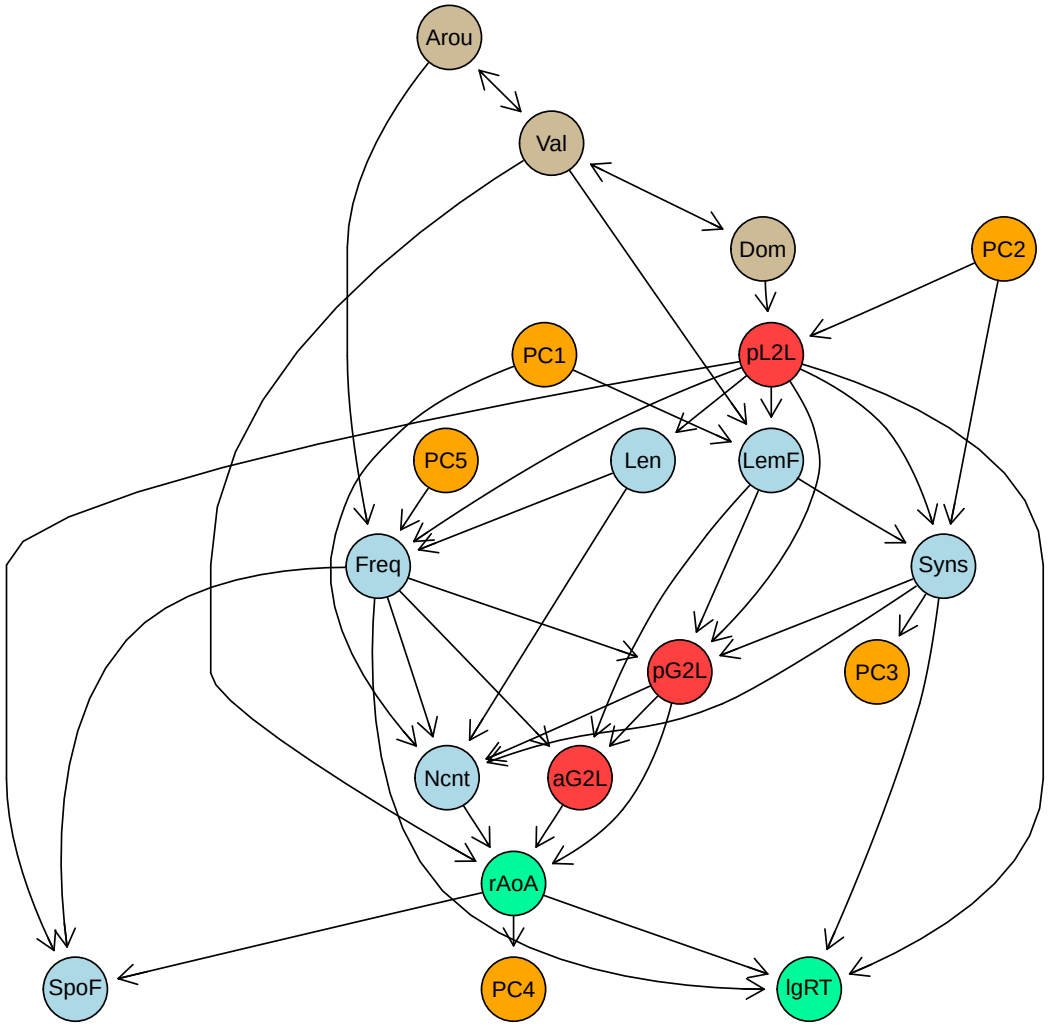


Figure 9: Essential graph. Emotion variables in gray, principal components in orange, lexical variables in blue, NDL measures in red, and age of acquisition and RT in green. `rAoA`: Rated Age of Acquisition; `Freq`: Subtitle Frequency (1100 million words); `LemF`: Celex Lemma Frequency; `SpoF`: Spoken Frequency BNC; `Len`: Word Length; `Ncnt`: Neighbor Count; `Syns`: Synset Count; `PC1`–`PC5`: principal components of pivot words; `Val`: Valence; `Arou`: Arousal; `Dom`: Dominance; `pL2L`: NDL L2L prior; `pG2L`: NDL G2L prior; `aG2L`: NDL G2L activation; `lgRT`: log-transformed RT.

A remarkable aspect of the hierarchy disovered by the hill-climbing algorithm is that semantic measures (the three emotion measures, in gray, and of PCs, in yellow, those that are predicting (PC1, PC2, PC5) rather than being predicted (PC3, PC4)) precede the lexical distributional measures (in blue), which in turn precede Age of Acquisition and RT. Of the NDL measures (in red), the one that estimates word-to-word co-occurrence priors (`pL2L`) is the more semantic one and appears higher up, whereas the two measures that gauge learning from orthography are predicted by the lexical distributional measures and themselves predict Age of Acquisition. Although we believe considerable caution is required, it is interesting that if in the spirit of Pearl (2009) this hierarchy is indeed

interpretable as mirroring causality, we have by and large a flow of causality from semantic measures via measures of form to behavioral measures (reaction time and age of acquisition ratings). This flow of causality in the model is reminiscent of the process from conceptualization to articulation in speech production (Levelt, 1989), and may indicate strong constraints on the joint density of the present variables originating from speech production.

According to this model, Reaction Time is "caused" by four variables. They are shorter

1. when words occur more often in print,
2. when words have more meanings or senses,
3. when words are more likely to follow many other words, and
4. when words are more like the simple words in use from early childhood.

We find RT at the bottom of the hierarchy, as befits a chronometric measure elicited in a meta-linguistic task. RT is conditionally independent of each of the other remaining fourteen variables.

Age of Acquisition is determined by four different random variables. Words that are judged to be acquired earlier in life are

1. happy words (with positive valence),
2. words that are easy to articulate (indexed through a high neighborhood density),
3. words that are well activated from the visual input (high NDL G2L activation), and
4. words that have high orthographic priors (high NDL G2L prior).

The last three variables indicate a strong bias from ease of visual perception when participants rate words presented to them in writing. Happy words that are easy to read in turn give rise to shorter reaction times. These happy words also have semantic vectors that are more similar to that of the expletive *fuck* and less similar to the semantic vectors of *daddy* and *play*, and they tend to be used more often in the spoken language.

The causes and consequences of the age of acquisition measure in our model are consistent with the conclusion reached previously: The predicivity of Age of Acquisition for reaction times is unlikely to arise as a consequence of the age at which a word is acquired. Like RT, age of acquisition is an experimental response variable at the bottom of the causal hierarchy, and not a causal factor at the top of the hierarchy.

The fact that Age of Acquisition is predictive for reaction times may be due to the very similar ways in which these ratings and reaction times are collected. In both cases, the response variable is a consensus measure resulting from an averaging process over data collected from many participants. Counts of frequency of occurrence, number of neighbors, and number of synsets are not consensus measures. They are 'one-shot' measures based on a particular sample of the language that is not calibrated in any way across many individuals' experience.

Four variables determine Subtitle Frequency.

1. The more exciting a word is (the greater its arousal value), the more likely it is that it is used in writing;
2. the more a word is semantically similar to *mum* (and *daddy*), and the less it is similar to *sex* (and *sexy*), the more it will be used in writing;
3. the shorter a word is, the greater its chances of appearing in print; and

4. the more often a word is predicted from other words (NDL L2L Prior), the more probable its appearance in writing is.

Unsurprisingly, Subtitle Frequency (1100 million words) predicts Spoken Frequency in the BNC (5 million words), neighborhood density, the NDL G2L measures, and reaction time. Frequency is conditionally independent of Age of Acquisition given the Neighbor Count, Valence, and the two NDL G2L measures.

The essential graph of Figure 9 elegantly clarifies that frequency of occurrence is not a primary causal factor. To the contrary, we find frequency in the center of the essential graph, where it is being shaped by the forces of meaning, constraints on production effort (word length), and co-occurrence (NDL L2L Prior). Thus, frequence of occurrence emerges as part of a complex dynamic system. It is crucial to keep this system in mind. Not doing so results in vastly inflating the importance of this variable. By way of example, the model fitted to the response latencies with all predictors achieved an adjusted R-squared of 0.416. When we regress frequency on all other predictors, and take residuals in order to obtain a measure of "pure frequency of occurrence" uncontaminated by other predictors, the amount of variance in the reaction times explained reduces to 0.02, less than 5% of what can be explained. Likewise, the sum of the amounts of variance that is uniquely explained by each of the individual predictors is no more than 0.098, which is less than 20% of the variance that can be captured by a systems approach. Crucially, the bulk of the variance in reaction times that can be explained comes from the system, from the interplay of many — individually imperfect, but jointly powerful — random variables (see also Baayen, 2011).

The position of word length in the essential graph, wedged in between the lexome-to-lexome prior and subtitle frequency, may be of some theoretical interest. According to information theory (Shannon, 1948), the linguistic code will be efficient only if more frequent words are shorter. Although frequency and length have to be in balance, it is not immediately clear whether the two are in a causal relation, and if so, in which direction causality flows.

On the one hand, length may have a causal influence on frequency. Mandelbrot (1953) used a Markov model to generate words from the transitional probabilities of letters. If it is assumed that the cost of coding a word is linear in its length in letters, then the Zipfian shape of the rank-frequency distribution follows straightforwardly. Furthermore, as pointed out by Nusbaum (1985), such a Markov model not only correctly predicts shorter words to be more frequent, it also correctly predicts that more frequent words will have denser similarity neighborhoods (see also Baayen, 2001). In addition, this model successfully predicts that shorter words will be in the language longer, simply because shorter words have higher probabilities of being generated. The history of Chinese is interesting in this respect. Ancient Chinese was predominantly a language with short, monosyllabic words. Over time, under the onomasiological pressure of having to discriminate between exponentially growing numbers of technical innovations, Chinese has become a language in which compounds dominate (Arcodia, 2007). From this Markovian perspective, it is the phonotactics and the ease of articulation which arguably determine frequency: It is the word forms that are easy to produce that will be used more frequently.

On the other hand, there may also be a causal influence of frequency on length. The forms of a language are under relentless pressure towards simplification. Above, we already mentioned that acoustically reduced forms are highly common in conversational speech. Likewise, names for new inventions typically start out long, as appropriate for forms just entering the vocabulary. But when an invention becomes popular and enjoys frequent use, then its name will typically be shortened, as has happened to *automobile* and *television* in English (Zipf, 1949). Here, then, we have a societal change, an invention becoming popular, leading to a reduction in the length of the invention's name, suggesting a causal flow from frequency to length.

When we now consider the essential graph, we find that the lexome-to-lexome prior is the causal factor for word length. This prior is measure of how well a lexome is available a-priori vis-a-vis the other lexomes in the language, independently of visual (or auditory) input. This prior reflects the importance of a lexome for communication in general, and therefore appears a likely candidate for driving shortening when a word has developed a probability of use that is greater than would be commensurable for its length. Conversely, the dependence of subtitle frequency on length may reflect the importance of ease of articulation for frequent use.

# 7 Summary and conclusions

To summarize the present explorations at the interface of corpus design, register variation, lexical statistics, and lexical processing, we note that, first of all, frequency of occurrence is part of an intricate system of correlations with both lexical distributional variables and measures assessing aspects of emotion, unfortunately all too often based on a fortituous combination of structural language properties and orthographic conventions.

Complexities are multiplied by writers writing and speakers speaking to be effective and well understood. As a consequence, which words they use with preference varies along with their communicative intentions. When addressing the young, they may opt for using short words requiring articulatory gestures that the young can approximate. When crafting a novel, writers will seek to make use of the full potential for expression that the language allows, including the use of specialized words that are 'just right'. When composing subtitles, the emotional goals of movies and the constraints of needing text that can be scanned rapidly give rise to yet another text register with its own specific characteristics, different from those of normal day-to-day conversation. Tweets come with similar constraints as subtitles for movies or TV, and we anticipate that the good predictivity observed by Gimenes and New (2015) and van Heuven et al. (2014) for counts from these registers is grounded in a configuration of distributional correlations similar to the one we observed for subtitles. Critically, the change from one text type to another does not result in just different frequency counts, but also changes the correlational structure of a wide range of other lexical properties. For the study of lexical processing, this state of affairs is exacerbated by the crudeness of our frequency counts, the absence of proper sense disambiguation and lemmatization, and decontextualization (note that in the present study, the only variables taking into account the immediate context of a word are the ones based on discriminative learning).

The 'audience design' of a given register may, or may not, dovetail well with the requirements of a specific psychometric task. For tasks requiring rapid visual uptake of isolated words, such as lexical decision and word naming, registers favoring simple and short words that are easy to read and easy to say will provide frequency counts that predict reaction times better, especially if they also have arousal and valence values that provide participants in a boring task with more stimulation (see also Wurm, Vakoch, Aycock & Childers, 2003; Wurm & Vakoch, 1996; Wurm, 2007; Kryuchkova, Tucker, Wurm & Baayen, 2012, for the potential evolutionary relevance of emotion in lexical processing). The registers of the language for the young, and the register for rapid visual scanning of subtitles for movies acted out in an unfamiliar language, although optimal for predicting reaction times to isolated words, may not be helpful for understanding the fine details of how people read morphologically complex words in prose, or how they understand each other in normal day-to-day conversation. With respect to subtitle frequencies, we note that as the cultural distance between the original language of a film (typically, English) and the language in which the subtitles are written increases, the discrepancy between the subtitle register and actual language experience will increase as well. This may explain why Pham (2014) observed, for Vietnamese, that frequencies

from written text explained 1% to 5% more of the variance in reaction times in visual lexical decision than frequencies from Vietnamese subtitles.

A generally unanticipated side-effect of selecting a register that optimizes the predictivity of frequency of occurrence for a specific kind of task is that the importance of other predictors is likely to be reduced, or even masked. We have seen this for subtitle frequencies. Because movies exploit emotions and hence overuse emotionally laden words in comparison to normal speech and writing, frequency of occurrence becomes highly confounded with valence, arousal and dominance (or danger and usefulness, see Wurm, 2007), making it more difficult to ascertain the importance of these other measures for lexical processing.

Measures based on the mathematical formalization of discrimination learning emerged as strong determinants (perhaps even causal factors) of age of acquisition and reaction time in a Bayesian graphical model. The variable importance of the discrimination measures a random forests predicting age of acquisition and reaction time were substantial. For age of acquisition, the grapheme-to-lexome prior was by far the most important predictor, followed at a distance by lemma frequency, reaction time, and the other two learning measures. For the reaction time, subtitle frequency was more important than the lexome-to-lexome prior, but here the grapheme-to-lexome prior emerged with greater variable importance than word length, spoken frequency, neighborhood density, and the three measures of emotion. The theory of discrimination learning also allowed us to derive further measures, based on semantic similarity to pivotal words for both early childhood and adulthood, that successfully predicted both age of acquisition and reaction time.

The NDL grapheme-to-lexome and lexome-to-lexome priors, which also capture aspects of words' contextual diversity and contextual distinctiveness, are part of a theory in which the effect of frequency of occurrence can be understood without having to assume counters in the head. Counters in the head, such as resting activation levels or activation thresholds in interactive activation and spreading activation models, and also the priors of Bayesian models (see, e.g., Norris, 2006; Norris & McQueen, 2008) are required by theories failing to reflect on the learning involved for discriminating between words. Once the importance of learning is taken seriously, the (Zipfian, see Good, 1953; MacArthur, 1957) frequencies with which events and objects are encountered will, in interaction with the properties of these events and objects as discriminated by the speaker, drive association strengths such that their joint effect on lexical processing will mirror, albeit ever partially, these frequencies. The discriminative learning process over a lexicon in which connection strengths from sublexical cues to lexical outcomes are constantly recalibrated explains why pure frequency of occurrence, decorrelated from other lexical properties, accounts for only a minute proportion of the variance in reaction times (see also Baayen, 2011). Although we realize this remains to be shown, we anticipate that a theory in which lexical availability is distributed across sparse discriminative cues in a network, instead of being discretized into a frequency counter in the head, has much to offer for our understanding of the highly diffuse patterns of breakdown of lexical processing under physiological insult.

A limitation of the present study is that it has considered frequency of occurrence only in relation to a few of the many response variables that inform about language processing. Effects of frequency and other variables considered in the present study vary substantially across tasks and modalities. Even within a single task such as silent reading, the effects of frequency measures may vary between first fixations, subsequent fixations, and total fixation duration (see, e.g., Kuperman, Schreuder, Bertram and Baayen, 2009; Miwa, Libben, Dijkstra and Baayen, 2014; Hendrix, 2015 for the silent reading of compounds in Dutch and Japanese, see also Kuperman, Drieghe, Keuleers and Brysbaert, 2013 for a comparison of lexical decision and eye-tracking data). Furthermore, effects may differ depending on whether a word is read in isolation, or in a sentence context (Luke and Christianson, 2011). To complicate matters further, there are substantial individual differences

in lexical processing (Kuperman and Van Dyke, 2011) that likely reflect to a considerable extent very different personal histories of language experience. This brings us to a fundamental problem for a general frequency 'norm' (such as proposed by Carroll, 1970 and independently nearly half a century later by van Heuven et al., 2014), namely that it presupposes an ideal native speaker whose language input is properly sampled by the corpus on which the normative counts are based. Unfortunately, how people use language varies substantially with social group (Labov, 1972) and their individual habits, important co-determinants of the kind of language registers they are exposed to and participate in. Whereas television broadcasts may be an important source of input for a bus driver, English novels are likely to be a dominant source of input for an undergraduate psychology student doing a minor in English literature.

Even if it were possible to compile corpora for specific social groups and derive frequency counts from such targeted corpora, the possibility remains that shorter words with strong sexual and emotional connotations are processed more quickly, *in relative independence* of how often these words appear in the actual input. As pointed out by Wurm, 2007, words the understanding of which is essential for survival may have an independent processing advantage (see also Thomas and LaBar, 2005 for the strong priming effect of taboo words). Interestingly, it has been argued that a subcortical pathway exists for the processing of emotion-rich stimuli (Phelps and LeDoux, 2005; LeDoux, 2007), enabling adequate and quick "fight or flight" responses, and consistent with this possibility, Scott, O'Donnell, Leuthold and Sereno, 2009 and Kryuchkova et al., 2012 observed temporaly early effects in the EEG signal of emotion in visual and auditory comprehension respectively. Furthermore, words with negative valence as well as words with high arousal tend to be remembered better (Kensinger and Corkin, 2003), suggesting a qualitative advantage for the learning of emotion words. If it is indeed the case that emotion words have an advantage both in learning and in processing, then the construction of frequency norms from samples of language use in registers that favor frequent use of emotion words will make it extremely difficult to disentangle the relative contributions of frequency and emotion. To break this circularity, it will be necessary to construct 'input' corpora that approximate, to the best of our knowledge, the experience of the subject subpopulation sampled for an experiment. We anticipate that the frequency counts based on such corpora (and likewise similarly measures based on discriminative learning) will explain less of the variance in measures such as lexical decision response times than a corpus of subtitles or a corpus of tweets. But exactly the difference in the predictions of 'input' corpora and subtitle or tweet corpora will be informative about the extent to which the learning of emotion words is advantaged beyond simple frequency of occurrence.

For the practical assessment of language impairments, the central point of the present study is the importance of the kind of language sampled by the corpus from which frequency counts have been extracted, vis-a-vis the cumulative language experience of a given speaker or a sample of speakers. Questions concerning the constraints with which writers have to work, the articulatory skills (or lack thereof) of addressees, the communicative and economic goals of the register, and how these factors interact with the language experience of a speaker at a given age, all have to be considered carefully for a proper evaluation of the role of frequency of occurrence in aphasia.

## References


Adelman, J., Brown, G. & Quesada, J. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814.

Arcodia, G. F. (2007). Chinese: a language of compound words? In *Selected proceedings of the 5th décembrettes: Morphology in Toulouse* (pp. 79–90).

Arnon, I. & Priva, U. C. (2013). More than words: the effect of multi-word frequency and constituency on phonetic duration. *Language and speech*, *56*(3), 349–371.

Arnon, I. & Snider, N. (2010). More than words: frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82.

Baayen, R. H. (1996). The effect of lexical specialization on the growth curve of the vocabulary. *Computational Linguistics*, *22*, 455–480.

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. (2011). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, *5*, 436–461.

Baayen, R. H. (2013). Multivariate Statistics. In R. Podesva & D. Sharma (Eds.), *Research Methods in Linguistics* (pp. 337–372). Cambridge: Cambridge University Press.

Baayen, R. H., Feldman, L. & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *53*, 496–512.

Baayen, R. H., Kuperman, V. & Bertram, R. (2010). Frequency effects in compound processing. In S. Scalise & I. Vogel (Eds.), *Compounding*. Amsterdam/Philadelphia: Benjamins.

Baayen, R. H., Milin, P., Filipović Durdević, D., Hendrix, P. & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*, 438–482. doi:10.1037/a0023851

Baayen, R. H. & Moscoso del Prado Martín, F. (2005). Semantic density and past-tense formation in three Germanic languages. *Language*, *81*, 666–698.

Baayen, R. H., Piepenbrock, R. & Gulikers, L. (1995). *The CELEX lexical database (cd-rom)*. University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Baayen, R. H., Shaoul, C., Willits, J. & Ramscar, M. (2015). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition, and Neuroscience*. doi:DOI:10.1080/23273798.2015.1065336

Baayen, R. H., Tomaschek, F., Gahl, S. & Ramscar, M. (2015). The Ecclesiastes principle in language change. In M. Hundt, S. Mollin & S. Pfenninger (Eds.), *The changing English language: Psycholinguistic perspectives* (to appear). Cambridge, UK: Cambridge University Press.

Baayen, R. H., Van Halteren, H. & Tweedie, F. (1996). Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, *11*, 121–131.

Baayen, R. H., Wurm, L. H. & Aycock, J. (2007). Lexical dynamics for low-frequency complex words. a regression study across tasks and modalities. *The Mental Lexicon*, *2*, 419–463.

Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D. & Yap, M. (2004). Visual word recognition for single-syllable words. *Journal of Experimental Psychology: General*, *133*, 283–316.

Bannard, C. & Matthews, D. (2008). Stored word sequences in language learning: the effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*, 241–248.

Barber, H., Vergara, M. & Carreiras, M. (2004). Syllable-frequency effects in visual word recognition: evidence from erps. *Neuroreport*, *15*(3), 545–548.

Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, *43*(3), 209–226.

Barredo, J., Öztekin, I. & Badre, D. (2013). Ventral fronto-temporal pathway supporting cognitive control of episodic memory retrieval. *Cerebral Cortex*, bht291.

Barry, C., Hirsh, K., Johnston, R. A. & Williams, C. (2001). Age of acquisition, word frequency, and the locus of repetition priming of picture naming. *Journal of Memory and Language*, *44*, 350–375.

Bauer, P. J. & Larkina, M. (2014). The onset of childhood amnesia in childhood: a prospective investigation of the course and determinants of forgetting of early-life events. *Memory*, *22*(8), 907–924.

Bekinschtein, P. & Weisstaub, N. (2014). Role of pfc during retrieval of recognition memory in rodents. *Journal of Physiology-Paris*, *108*(4), 252–255.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1989). A typology of english texts. *Linguistics*, *27*, 3–43.

Bowers, J., Davis, C. & Hanley, D. (2005). Automatic semantic activation of embedded words: Is there a "hat" in "that"? *Journal of Memory and Language*, *52*, 131–143. doi:10.1016/j.jml.2004.09.003

Brants, T. & Franz, A. (2006). *Web 1t 5-gram version 1*. Philadelphia: Linguistic Data Consortium.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brysbaert, M. (1996). Word frequency affects naming latency in Dutch with age-of-acquisition controlled. *European Journal of Cognitive Psychology*, *8*, 185–193.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J. & Böhl, A. (2015). The word frequency effect. *Experimental Psychology*.

Brysbaert, M. & Ellis, A. W. (2015). Aphasia and age-of-acquisition: are early-learned words more resilient? *Aphasiology*, *in press*.

Brysbaert, M., Keuleers, E. & New, B. (2011). Assessing the usefulness of google books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology*, *2*.

Brysbaert, M., Lange, M. & Van Wijnendaele, I. (2000). The effects of age-of-acquisition and frequency-of-occurence in visual word recognition: further evidence from the Dutch language. *European Journal of Cognitive Psychology*, *12*, 65–85.

Brysbaert, M. & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, *41*(4), 977–990.

Burnard, L. (1995). *Users guide for the British National Corpus*. Oxford university computing service: British National Corpus consortium.

Burrows, J. F. (1987). Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, *2*, 61–70.

Burrows, J. F. (1992). Computers and the study of literature. In C. S. Butler (Ed.), *Computers and written texts* (pp. 167–204). Oxford: Blackwell.

Carreiras, M., Alvarez, C. J. & de Vega, M. (1993). Syllable frequency and visual word recognition in Spanish. *Journal of Memory and Language*, *32*(6), 766–780.

Carroll, J. B. (1970). An alternative to Juilland's usage coefficient for lexical frequencies, and a proposal for a standard frequency index (sfi). *Computer Studies in the Humanities and Verbal Behavior*, *3*, 61–65.

Carroll, J. B. & White, M. N. (1973a). Age of acquisition norms for 220 picturable nouns. *Journal of Verbal Learning and Verbal Behavior*, *12*, 563–576.

Carroll, J. B. & White, M. N. (1973b). Word frequency and age of acquisition as determiners of picture-naming latency. *Quaterly Journal of Experimental Psychology*, *25*, 85–95.

Cholin, J., Schiller, N. O. & Levelt, W. (2004). The preparation of syllables in speech production. *Journal of Memory and Language*, *20*(50), 47–61.

Chugani, H. T. (1996). Neuroimaging of developmental nonlinearity and developmental pathologies. *Developmental neuroimaging: Mapping the development of brain and behavior*, 187–195.

Church, K. & Gale, W. (1995). Inverse document frequency (IDF): a measure of deviations from Poisson. In D. Yarowsky & K. Church (Eds.), *Third workshop on very large corpora* (pp. 121–130). ACL. MIT.

Coltheart, M., Rastle, K., Perry, C., Langdon, R. & Ziegler, J. (2001). The DRC model: A model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–258.

Conway, M. A. (2005). Memory and the self. *Journal of Memory and Language*, *53*(4), 594–628.

Cortese, M. J. & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: an analysis of 2342 words. *Quarterly Journal of Experimental Psychology*, *60*, 1072–1082.

Davies, M. (2010). The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing*, *25*(4), 447–464.

De Saussure, F. (1966). *Course in general linguistics*. New York: McGraw.

Dijkstra, A. & Van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: from identification to decision. *Bilingualism: Language and Cognition*, *5*, 175–197.

Ellis, A. W. & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflects loss of plasticity in maturing systems: Insights from connectionist networks. *JEP:LMC*, *26*, 1103–1123.

Ernestus, M., Baayen, R. H. & Schreuder, R. (2002). The recognition of reduced word forms. *Brain and Language*, *81*, 162–173. doi:10.1006/brln.2001.2514

Fernald, A., Marchman, V. A. & Weisleder, A. (2013). Ses differences in language processing skill and vocabulary are evident at 18 months. *Developmental science*, *16*(2), 234–248.

Fletcher, P., Shallice, T., CD, F., SJ, F. & Dolan, R. (1998). The functional roles of prefrontal cortex in episodic memory. ii. retrieval. *Brain*, *121*, 1249–1256.

Fletcher, P., Shallice, T. & Dolan, R. (1998). The functional roles of prefrontal cortex in episodic memory. i. encoding. *Brain*, *121*(7), 1239–1248.

Freud, S. (1905/1953). Childhood and concealing memories. In A. A. Brill (Ed.), *The basic writings of Sigmund Freud*. New York: The Modern Library.

Gahl, S. (2008). Time and thyme are not homophones: the effect of lemma frequency on word durations in spontaneous speech. *Language*, *84*(3), 474–496.

Gahl, S., Yao, Y. & Johnson, K. (2012). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806.

Gardner, M. K., Rothkopf, E. Z., Lapan, R. & Lafferty, T. (1987). The word frequency effect in lexical decision: finding a frequency-based component. *Memory and Cognition*, *15*, 24–28.

Gentry, J., Long, L., Gentleman, R., Falcon, S., Hahne, F., Sarkar, D. & Hansen, K. D. (2015). *Rgraphviz: provides plotting capabilities for r graph objects*. R package version 2.4.0.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*, 256–281.

Ghetti, S. & Bunge, S. A. (2012). Neural changes underlying the development of episodic memory during middle childhood. *Developmental cognitive neuroscience*, *2*(4), 381–395.

Ghyselinck, M., Lewis, M. B. & Brysbaert, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: a review of the literature and a new multi-task investigation. *Acta Psychologica*, *115*(1), 43–67.

Gilhooly, K. J. & Gilhooly, M. L. M. (1980). The validity of age-of-acquisition ratings. *British Journal of Psychology*, *71*, 105–110.

Gimenes, M. & New, B. (2015). Worldlex: twitter and blog word frequencies for 66 languages. *Behavior research methods*, in press.

Glanzer, M. & Bowles, N. (1976). Analysis of the word frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 21–31.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*, 237–264.

Goodman, J. C., Dale, P. S. & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of child language*, *35*(3), 515.

Gordon, E., Maclagan, M. & Hay, J. (2007). The ONZE Corpus. In J. Beal, K. Corrigan & H. Moisl (Eds.), *Models and methods in handling of unconventional digital corpora* (Vol. 2, pp. 82–104). Palgrave.

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language*, *29*, 228–244.

Grainger, J. & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: a multiple read-out model. *Psychological Review*, *103*, 518–565.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International journal of corpus linguistics*, *13*(4), 403–437.

Gries, S. T. (2010). Dispersions and adjusted frequencies in corpora: further explorations. *Corpus linguistic applications: current studies, new directions*, 197–212.

Griffin, Z. & Bock, K. (1998). Constraint, Word Frequency, and the Relationship between Lexical Processing Levels in Spoken Word Production. *Journal of Memory and Language*, *38*, 313–338.

Habermas, T. & Bluck, S. (2000). Getting a life: the emergence of the life story in adolescence. *Psychological bulletin*, *126*(5), 748.

Halgren, E., Dhond, R. P., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J. D. & Dale, A. M. (2002). N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. *Neuroimage*, *17*(3), 1101–1116.

Halteren, H. v., Baayen, R. H., Tweedie, F., Haverkort, M. & Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, *12*, 65–77.

Harm, M. W. & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*, 491–528.

Heister, J. & Kliegl, R. (2012). Comparing word frequencies from different german text corpora. *Lexical Resources in Psycholinguistic Research*, *3*, 27–44.

Hendriks, P., Englert, C., Wubs, E. & Hoeks, J. (2008). Age differences in adults' use of referring expressions. *Journal of Logic, Language and Information*, *17*(4), 443–466.

Hendrix, P. (2015). *Experimental explorations of a discrimination learning approach to language processing* (Doctoral dissertation, University of Tübingen).

Henri, V. & Henri, C. (1895). On our earliest recollections of childhood. *Psychological Review*, *2*, 215–216.

Hofstetter, S., Tavor, I., Moryosef, S. T. & Assaf, Y. (2013). Short-term learning induces white matter plasticity in the fornix. *The Journal of Neuroscience*, *33*(31), 12844–12850.

Højsgaard, S., Edwards, D. & Lauritzen, S. (2012). *Graphical models with r*. Springer Science & Business Media.

Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A. & Van Der Laan, M. (2006). Survival ensembles. *Biostatistics*, *7*, 355–373.

Hurtado, N., Marchman, V. A. & Fernald, A. (2008). Does input influence uptake? links between maternal talk, processing speed and vocabulary size in spanish-learning children. *Developmental science*, *11*(6), F31–F39.

Jacques, P. L. S., Kragel, P. A. & Rubin, D. C. (2011). Dynamic neural networks supporting memory retrieval. *Neuroimage*, *57*(2), 608–616.

Jescheniak, J. D. & Levelt, W. (1994). Word frequency effects in speech production: retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*(4), 824–843.

Johnson, K. (2004). Massive reduction in conversational American English. In *Spontaneous speech: data and analysis. proceedings of the 1st session of the 10th international symposium* (pp. 29–54). The National International Institute for Japanese Language. Tokyo, Japan.

Johnson, N. L. & Kotz, S. (1977). *Urn models and their application. an approach to modern discrete probability theory*. New York: John Wiley & Sons.

Jones, M., Curran, T., Mozer, M. C. & Wilder, M. H. (2013). Sequential effects in response time reveal learning mechanisms and event representations. *Psychological review*, *120*(3), 628.

Kensinger, E. A. & Corkin, S. (2003). Memory enhancement for emotional words: are emotional words more vividly remembered than neutral words? *Memory & cognition*, *31*(8), 1169–1180.

Keuleers, E., Lacey, P., Rastle, K. & Brysbaert, M. (2012). The british lexicon project: lexical decision data for 28,730 monosyllabic and disyllabic english words. *Behavior Research Methods*, *44*(1), 287–304.

Keuleers, E., Stevens, M., Mandera, P. & Brysbaert, M. (2015). Word knowledge in the crowd: measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology*, 1–62.

Keune, K., Ernestus, M., Van Hout, R. & Baayen, R. H. (2005). Social, geographical, and register variation in Dutch: From written 'mogelijk' to spoken 'mok'. *Corpus Linguistics and Linguistic Theory*, *1*, 183–223.

Koesling, K., Kunter, G., Baayen, R. H. & Plag, I. (2012). Prominence in triconstituent compounds: pitch contours and linguistic theory. *Language and Speech*, *56*(4), 529–554.

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Kryuchkova, T., Tucker, B. V., Wurm, L. & Baayen, R. H. (2012). Danger and usefulness in auditory lexical processing: evidence from electroencephalography. *Brain and Language*, *122*, 81–91.

Kučera, H. & Francis, W. N. (1967). *Computational analysis of present-day american english*. Providence, RI: Brown University Press.

Kuperman, V., Schreuder, R., Bertram, R. & Baayen, R. H. (2009). Reading of multimorphemic Dutch compounds: Towards a multiple route model of lexical processing. *Journal of Experimental Psychology: HPP*, *35*, 876–895.

Kuperman, V. & Van Dyke, J. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of memory and language*, *65*(1), 42–73.

Kuperman, V. & Bertram, R. (2013). Moving spaces: spelling alternation in english noun-noun compounds. *Language and Cognitive Processes*, *28*(7), 939–966.

Kuperman, V., Drieghe, D., Keuleers, E. & Brysbaert, M. (2013). How strongly do word reading times and lexical decision times correlate? combining data from eye movement corpora and megastudies. *The Quarterly Journal of Experimental Psychology*, *66*(3), 563–580.

Kuperman, V., Stadthagen-Gonzalez, H. & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, *44*(4), 978–990.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Landauer, T. & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Lauritzen, S. L. (1996). *Graphical models*. Oxford, UK: Clarendon Press.

LeDoux, J. (2007). The amygdala. *Current Biology*, *17*(20), 868–874.

Levelt, W. (1989). *Speaking. from intention to articulation*. Cambridge, Mass.: The MIT Press.

Levelt, W., Roelofs, A. & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–38.

Luciana, M. & Nelson, C. A. (1998). The functional emergence of prefrontally-guided working memory systems in four-to eight-year-old children. *Neuropsychologia*, *36*(3), 273–293.

Luke, S. G. & Christianson, K. (2011). Stem and whole-word frequency effects in the processing of inflected verbs in and out of a sentence context. *Language and Cognitive Processes*, *26*(8), 1173–1192.

MacArthur, R. H. (1957). On the relative abundance of bird species. *Proceedings of the National Academy of Sciences of the United States of America*, *43*(3), 293.

MacWhinney, B. (2000). The childes project. *Tools for Analyzing Talk. Part*, *1*.

Mandelbrot, B. (1953). An information theory of the statistical structure of language. In W. E. Jackson (Ed.), *Communication theory* (pp. 503–512). New York: Academic Press.

Mazzoni, G., Scoboria, A. & Harvey, L. (2010). Nonbelieved memories. *Psychological Science*.

McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: part i. an account of the basic findings. *Psychological Review*, *88*, 375–407.

McDonald, S. & Shillcock, R. (2001). Rethinking the word frequency effect: the neglected role of distributional information in lexical processing. *Language and Speech*, *44*, 295–323.

McDonald, S. & Ramscar, M. (2001). Testing the distributional hypothesis: the influence of context judgements of semantic similarity. In *Proceedings of the 23rd annual conference of the cognitive science society*.

McRae, K., Jared, D. & Seidenberg, M. S. (1990). On the roles of frequency and lexical access in word naming. *Journal of Memory and Language*, *29*, 43–65.

Meunier, F. & Segui, J. (1999). Frequency effects in auditory word recognition: the case of suffixed words. *Journal of Memory and Language*, *41*, 327–344.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Milin, P., Filipović Durdević, D. & Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*, 50–64.

Milin, P., Ramscar, M., Baayen, R. H. & Feldman, L. B. (2015). Cornering segmentation: the perspective from discrimination learning. *Manuscript submitted for publication, University of Tübingen*.

Miller, G. A. (1990). Wordnet: an on-line lexical database. *International Journal of Lexicography*, *3*, 235–312.

Miwa, K., Libben, G., Dijkstra, T. & Baayen, R. H. (2014). The time-course of lexical activation in japanese morphographic word recognition: evidence for a character-driven processing model. *The Quarterly Journal of Experimental Psychology*, *67*(1), 79–113.

Morrison, C. M., Chappell, T. D. & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *The Quarterly Journal of Experimental Psychology: Section A*, *50*(3), 528–559.

Morrison, C. M. & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *JEP:LMC*, *21*, 116–133.

New, B., Brysbaert, M., Veronis, J. & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(04), 661–677.

Nguyen, T. G. (2011). *Van de "tu" trong tieng viet [the issues of "word" in vietnamese]*. Ha Noi: Nha xuat ban Giao Duc.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*(2), 327–357.

Norris, D. & McQueen, J. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. doi:10.1037/0033-295X.115.2.357

Nusbaum, H. C. (1985). *A stochastic account of the relationship between lexical density and word frequency*. Indiana University. Research on Speech Perception, Progress Report #11.

Oostdijk, N. (2002). The design of the Spoken Dutch Corpus. In P. Peters, P. Collins & A. Smith (Eds.), *New frontiers of corpus research* (pp. 105–112). Amsterdam: Rodopi.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Peters, G. J., David, C. N., Marcus, M. D. & Smith, D. M. (2013). The medial prefrontal cortex is critical for memory retrieval and resolving interference. *Learning & Memory*, *20*(4), 201–209.

Pham, H. (2014). *Visual processing of vietnamese compound words: a multivariate analysis of using corpus linguistic and psycholinguistic paradigms*. PhD dissertation, University of Alberta, Edmonton.

Pham, H. & Baayen, R. H. (2015). Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition, and Neuroscience*, to appear. doi:DOI:10.1080/23273798.2015.1054844

Phelps, E. & LeDoux, J. (2005). Contributions of the amygdala to emotion processing: from animal models to human behavior. *Neuron*, *48*(2), 175–187.

Pitt, M., Johnson, K., Hume, E., Kiesling, S. & Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95.

Pluymaekers, M., Ernestus, M. & Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustical Society of America*, *118*, 2561–2569.

R Development Core Team. (2015). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Ramscar, M., Dye, M. & McCauley, S. M. (2013). Error and expectation in language learning: the curious absence of mouses in adult speech. *Language*, *89*(4), 760–793. doi:10.1353/lan.2013.0068

Ramscar, M. & Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends In Cognitive Science*, *11*(7), 274–279.

Ramscar, M., Hendrix, P., Love, B. & Baayen, R. H. (2013). Learning is not decline: The mental lexicon as a window into cognition across the lifespan. *The Mental Lexicon*, *8*, 450–481. doi:10.1075/ml.8.3.08ram

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P. & Baayen, R. H. (2014). Nonlinear dynamics of lifelong learning: the myth of cognitive decline. *Topics in Cognitive Science*, *6*, 5–42. doi:10.1111/tops.12078

Ramscar, M., Smith, A., Dye, M., Futrell, R., Hendrix, P., Baayen, R. H. & Starr, R. (2013). The 'universal' structure of name grammars and the impact of social engineering on the evolution of natural information systems. In *Proceedings of the 35th meeting of the cognitive science society*. Berlin, Germany.

Ramscar, M., Yarlett, D., Dye, M., Denny, K. & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957. doi:10.1111/j.1551-6709.2009.01092.x

Rayner, K. & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*(3), 191–201.

Rescorla, R. A. & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.),

*Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton Century Crofts.

Sadat, J., Martin, C. D., Costa, A., Alario, F. et al. (2014). Reconciling phonological neighborhood effects in speech production through single trial analysis. *Cognitive psychology*, *68*, 33–58.

Scarborough, D. L., Cortese, C. & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 1–17.

Scarf, D., Gross, J., Colombo, M. & Hayne, H. (2013). To have and to hold: episodic memory in 3-and 4-year-old children. *Developmental Psychobiology*, *55*(2), 125–132.

Scott, G. G., O'Donnell, P. J., Leuthold, H. & Sereno, S. C. (2009). Early emotion word processing: evidence from event-related potentials. *Biological psychology*, *80*(1), 95–104.

Scutari, M. (2010). Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, *35*(3), 1–22. Retrieved from http://www.jstatsoft.org/v35/i03/

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Shaoul, C., Baayen, R. H. & Westbury, C. F. (2015). N-gram probability effects in a cloze task. *The Mental Lexicon*, *9*(3), 437–472.

Shaoul, C., Westbury, C. F. & Baayen, R. H. (2013). The subjective frequency of word n-grams. *Psihologija*, *46*(4), 497–537.

Shaoul, C. & Westbury, C. (2010). Exploring lexical co-occurrence space using hidex. *Behavior Research Methods*, *42*(2), 393–413.

Shaoul, C., Willits, J., Ramscar, M., Milin, P. & Baayen, R. H. (2015). A discrimination-driven model for the acquisition of lexical knowledge in auditory comprehension. *under revision*.

Somerville, L. H. & Casey, B. (2010). Developmental neurobiology of cognitive control and motivational systems. *Current opinion in neurobiology*, *20*(2), 236–241.

Stadthagen-Gonzalez, H. & Davis, C. J. (2006). The bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, *38*(4), 598–605.

Stemberger, J. P. & MacWhinney, B. (1986). Frequency and the lexical storage of regularly inflected forms. *Memory and Cognition*, *14*, 17–26.

Stolarova, M., Brielmann, A. A., Wolf, C., Rinker, T. & Baayen, R. H. (2015). Assessing gender, bilingualism, type and duration of early care as potential predictors of vocabulary size and composition at two years of age. *manuscript submitted for publication*.

Strange, D. & Hayne, H. (2013). The devil is in the detail: children's recollection of details about their prior experiences. *Memory*, *21*(4), 431–443.

Strobl, C., Malley, J. & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological methods*, *14*(4), 26.

Taft, M. (1994). Interactive-activation as a framework for understanding morphological processing. *LCP*, *9*(3), 271–294. doi:10.1080/01690969408402120

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology*, *57A*, 745–765.

Tagliamonte, S. & Baayen, R. H. (2012). Models, forests and trees of york english: was/were variation as a case study for statistical practice. *Language Variation and Change*, *24*, 135–178.

Thomas, L. & LaBar, K. (2005). Emotional arousal enhances word repetition priming. *Cognition & Emotion*, *19*(7), 1027–1047.

Thompson-Schill, S. L., Ramscar, M. & Chrysikou, E. G. (2009). Cognition without control when a little frontal lobe goes a long way. *Current Directions in Psychological Science*, *18*(5), 259–263.

Thorndike, E. L. & Lorge, I. (1944). *A teacher's word book of 30.000 words*. New York: Columbia University Press.

Tomaschek, F., Tucker, B. V., Wieling, M. & Baayen, R. H. (2014). Vowel articulation affected by word frequency. In *Proceedings of 10th ISSP, cologne* (pp. 429–432).

Tomaschek, F., Wieling, M., Arnold, D. & Baayen, R. H. (2013). Frequency effects on the articulation of German i and u: evidence from articulography. In *Proceedings of interspeech, lyon* (pp. 1302–1306).

Tomasello, M. (2009). *Constructing a language: a usage-based theory of language acquisition*. Harvard University Press.

Tremblay, A. & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on Formulaic Language: Acquisition and communication* (pp. 151–173). London: The Continuum International Publishing Group.

Tremblay, A., Derwing, B., Libben, G. & Westbury, C. (2011). Processing advantages of lexical bundles: evidence from self-paced reading and sentence recall tasks. *Language Learning*.

Tsamardinos, I., Aliferis, C. F. & Statnikov, A. R. (2003). Algorithms for large scale markov blanket discovery. In *FLAIRS conference* (Vol. 2).

Van Abbema, D. & Bauer, P. (2005). Autobiographical memory in middle childhood: recollections of the recent and distant past. *Memory*, *13*(8), 829–845.

van Heuven, W. J., Mandera, P., Keuleers, E. & Brysbaert, M. (2014). Subtlex-uk: a new and improved word frequency database for british english. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.

Warriner, A. B., Kuperman, V. & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207.

Weisleder, A. & Fernald, A. (2013). Talking to children matters early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–2152.

Wells, C., Morrison, C. M. & Conway, M. A. (2014). Adult recollections of childhood memories: what details can be recalled? *The Quarterly Journal of Experimental Psychology*, *67*(7), 1249–1261.

Westbury, C. (2014). You can't drink a word: lexical and individual emotionality affect subjective familiarity judgments. *Journal of Psycholinguistic Research*, *43*(5), 631–649.

Wingfield, A. (1968). Effect of frequency on identification and naming of objects. *American Journal of Psychology*, *81*, 226–234.

Wood, S. N. (2006). *Generalized Additive Models*. New York: Chapman & Hall/CRC.

Wright, C. (1979). Duration differences between rare and common words and their implications for the interpretation of word frequency effects. *Memory & Cognition*, *7(6)*, 411–419.

Wurm, L. H. (2007). Danger and usefulness: an alternative framework for understanding rapid evaluation effects in perception? *Psychonomic Bulletin & Review*, *14*, 1218–1225.

Wurm, L. H. & Vakoch, D. A. (1996). Dimensions of speech perception: semantic associations in the affective lexicon. *Cognition and Emotion*, *10*, 409–423.

Wurm, L. H., Vakoch, D. A., Aycock, J. & Childers, R. R. (2003). Semantic effects in lexical access: evidence from single-word naming. *Cognition and emotion*, *17*, 547–565.

Xiao, R. (2008). Well-known and influential corpora. In A. Lüdeling & M. Kyto (Eds.), *Corpus linguistics: an international handbook [volume 1]* (pp. 383–457). Berlin: Mouton de Gruyter.

Young, M. P. & Rugg, M. D. (1992). Word frequency and multiple repetition as determinants of the modulation of event-related potentials in a semantic classification task. *Psychophysiology*, *29*(6), 664–676.

Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.

Zhang, H., Huang, C.-R. & Yu, S. (2004). Distributional consistency: as a general method for defining a core lexicon. In *Lrec*.

Zhao, Y. & Jurafsky, D. (2009). The effect of lexical frequency and Lombard reflex on tone hyperarticulation. *Journal of Phonetics*, *37*(2), 231–247.

Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology*, *15*, 1–95.

Zipf, G. K. (1935). *The psycho-biology of language*. Boston: Houghton Mifflin.

Zipf, G. K. (1949). *Human behavior and the principle of the least effort. an introduction to human ecology*. New York: Hafner.

# A Appendix

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| Intercept | 6.8098 | 0.0438 | 155.4875 | $< 0.0001$ |
| B. smooth terms | edf | Ref.df | F-value | p-value |
| s(Subtitle Frequency) | 1.0001 | 1.0002 | 15.1488 | 0.0001 |
| s(Spoken Frequency) BNC | 4.6219 | 5.5144 | 10.9470 | $< 0.0001$ |
| s(Word Length) | 1.9180 | 1.9908 | 5.9249 | 0.0028 |
| te(Lemma Freq. and PC1) | 4.8135 | 5.7185 | 34.9774 | $< 0.0001$ |
| s(PC2) | 3.7381 | 4.7083 | 23.1244 | $< 0.0001$ |
| s(PC3) | 3.2427 | 4.1108 | 2.5433 | 0.0365 |
| s(PC4) | 1.5637 | 1.9653 | 19.5751 | $< 0.0001$ |
| s(PC5) | 3.1936 | 4.0600 | 9.5615 | $< 0.0001$ |
| s(Arousal) | 3.6020 | 4.5219 | 3.0776 | 0.0120 |
| s(Valence) | 4.3364 | 5.3751 | 18.3926 | $< 0.0001$ |
| s(Dominance) | 1.6646 | 2.1111 | 0.7791 | 0.4619 |
| s(N count) | 1.7234 | 2.1451 | 38.5747 | $< 0.0001$ |
| s(Synset Count) | 2.5471 | 3.2583 | 6.6177 | 0.0001 |

Table A.1: Summary of the partial effects in the generalized additive model fitted to rated age of acquisition. s(): thin plate regression spline; te(): tensor product smooth.

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| Intercept | 6.3593 | 0.0012 | 5130.7907 | $< 0.0001$ |
| B. smooth terms | edf | Ref.df | F-value | p-value |
| s(AoA) | 3.4405 | 4.3250 | 37.0430 | $< 0.0001$ |
| s(Lemma Frequency) | 5.4443 | 6.6928 | 23.6113 | $< 0.0001$ |
| s(Word Length) | 1.0001 | 1.0001 | 30.1695 | $< 0.0001$ |
| s(PC2) | 1.0001 | 1.0001 | 19.8417 | $< 0.0001$ |
| s(PC5) | 1.0000 | 1.0001 | 19.0906 | $< 0.0001$ |
| s(N count) | 2.3727 | 2.9344 | 13.8578 | $< 0.0001$ |
| s(Synset Count) | 1.0000 | 1.0000 | 18.3708 | $< 0.0001$ |
| ti(PC1) | 2.7483 | 3.2839 | 7.2606 | $< 0.0001$ |
| ti(Subtitle Frequency) | 2.6329 | 3.1673 | 208.4844 | $< 0.0001$ |
| ti(tensor product Subt. Freq. and PC1) | 2.1815 | 2.6800 | 4.0930 | 0.0095 |

Table A.2: Summary of the partial effects in the generalized additive model fitted to log RT. s(): thin plate regression spline, ti(): partial effect in a decompositional tensor product smooth.