Proceedings of CLS 48 (2012), 447-461
© Chicago Linguistic Society 2014. All rights reserved.


# Revisiting frequency and storage in morphological processing*

Constantine Lignos and Kyle Gorman
University of Pennsylvania

## 1 Introduction

The balance between storage and computation of complex words is a major point of departure both for theories of lexical representation (e.g., Goldberg 2006, Halle & Marantz 1993, Jackendoff 1975) and processing (e.g., Baayen *et al.* 1997, Butterworth 1983, Taft 2004). The atoms of lexical memory that are implicated in lexical processing experiments—be they whole words, roots and affixes, or some combination thereof—must ultimately coincide with the units of morphological theory if the latter are to be theories of the mental lexicon.

One clue into the architecture of lexical memory comes from the presence of robust *frequency effects* in lexical decision tasks, in which subjects judge whether a written or spoken stimulus is a real word and processing complexity is measured by reaction time. The recognition of complex words is facilitated both by whole word frequency (as measured from some representative corpus) as well as *base* (also called *cluster* or *root*) *frequency*, the summed frequency of all words sharing the same root. Frequency effects have long been known to account for a large amount of variance in lexical decision latencies (e.g., Howes & Solomon 1951).

The existence of base frequency effects suggest that the roots posited by morphologists are stored in memory and are integral to lexical access. Many models of lexical processing, however, downplay the role of roots and of base frequency. Proponents of these *dual-route* models (e.g., Baayen *et al.* 1997, Baayen & Schreuder 1999, Caramazza *et al.* 1988) argue that complex words are also stored in memory and emphasize the role of *whole word frequency* in word recognition. Some studies (e.g., Sereno & Jongman 1997) deny base frequency effects altogether.

In this study, we investigate whole word frequency effects for English regularly inflected words reported by Alegre & Gordon (1999), Baayen *et al.* (2007), New *et al.* (2004), and Sereno & Jongman (1997). We model a very large database of visual lexical decision latencies with mixed effects regression, using residualization to control for confounding predictors of reaction time. The result is a more nuanced view of base and whole word frequency effects and the storage-computation debate.

After reviewing previous experimental work on complex word recognition (§2), we describe the data set and construct a statistical model of lexical decision latencies for regularly inflected complex words in English (§3). We use this model to investigate word and base frequency effects (§4) and conclude that the whole word frequency effects do not argue in favor of whole word storage; in fact, they are inconsistent with the predictions of dual-route processing models (§5–6).

*This research was supported by an NSF-IGERT grant to the University of Pennsylvania Institute for Research in Cognitive Science. Thanks to David Balota and Melvin Yap for access to the ELP trial-level data, and to David Embick, Charles Yang, and the CLS 48 audience for helpful comments.

## 2 Related work

The processing literature is replete with evidence that the units of lexical access may be smaller than words. Lexical decision experiments indicate that words (and non-word foils) are decomposed into constituents—which resemble the roots and affixes of morphological theory—at an early stage of processing, and that the products of decomposition are the primary units of lexical access. However, other experiments suggest that decomposition may not be the only mechanism for word recognition.

### 2.1 Evidence for decomposition

Non-word lexical decision latencies are one source of evidence for decomposition. Seminal studies by Taft & Forster (1975, 1976), Taft *et al.* (1986), and Caramazza *et al.* (1988) find that non-words like **re-sert*, which appear to be morphologically complex, take longer to reject that non-words which lack apparent morphological structure such as **refant*). Apparent morphological complexity appears to result in *false decomposition*, leading the processing system down a garden path.

Studies which influence lexical decision latencies with primes also provide evidence for decomposition. There is a rich body of literature attesting to facilitation of word recognition when prime and target are inflectional variants of the same root (e.g., Feldman & Fowler 1987, Kempley & Morton 1982, Marslen-Wilson *et al.* 1993, Meunier & Marslen-Wilson 2004, Murrell & Morton 1974, Orsolini & Marslen-Wilson 1997, Sonnenstuhl *et al.* 1999, Stanners *et al.* 1979a, Stockall & Marantz 2006). Some studies report that the facilitatory effect of inflectional priming is as large as identity priming (e.g., Fowler *et al.* 1985, Napps 1989). This morphological priming effect cannot be attributed to mere orthographic or semantic similarity between prime and target, as experiments with very brief prime exposure (i.e., below the threshold for conscious visual recognition) find that orthographic and semantic similarity do not prime (e.g., Marslen-Wilson *et al.* 2008, Rastle *et al.* 2000) or even inhibit target recognition (e.g., Baayen *et al.* 2007, Drews & Zwitserlood 1995, Grainger *et al.* 1991, Henderson *et al.* 1984).

### 2.2 Whole-word access

The presence of inflectional priming and false decomposition effects rules out a naïve view of word recognition (e.g., Butterworth 1983) as consisting of little more than a search for whole words in lexical memory.[1] However, it is not clear that decomposition is the only mechanism for complex word recognition. In particular, whole word frequency effects in lexical decision tasks are cited as evidence for whole word storage. A number of studies investigating recognition of regularly inflected words (e.g., Baayen *et al.* 1997, Burani *et al.* 1984, Schriefers *et al.* 1992) find that when whole word frequency is held constant, base frequency facilitates word recognition, and when base frequency is held constant, there is a facilitatory effect of whole word frequency.

[1] Chan (2008) and Hankamer (1992) reach the same conclusion by a different method: they argue that the number of unique words in morphologically rich languages is so great—in some languages, infinite—that no speaker could hope to encounter (or store) more than a tiny portion of these words in a lifetime. But, since speakers are capable of understanding and producing words they have never before heard, some system for recognizing (and producing) complex words is needed.

There is some disagreement as to whether English regularly inflected words, particularly those of high frequency, exhibit root frequency effects. Alegre & Gordon (1999), New *et al.* (2004), and Taft (1979, 2004) find robust base frequency effects, but Sereno & Jongman (1997) and Baayen *et al.* (2007) do not. Alegre & Gordon (1999) find both whole word and base frequency effects, but report that whole word frequency effects are absent among items with low whole word frequency. They argue for a dual-route processing model in which high frequency words are stored and accessed as undecomposed wholes, but decomposition is used for low frequency words. However, Baayen *et al.* report that even low frequency words exhibit whole word frequency effects but only marginal root frequency effects.

### 2.3 Diagnosis of storage

Frequency effects are simply correlations between processing time and frequency of an item or its components. Researchers have generally assumed that these correlations reflect the granularity of lexical memory. For instance, if whole word storage were the only atom of the mental lexicon, one would expect that whole words, but not subcomponents like roots, would exhibit frequency effects. While this naïve assumption is a useful heuristic, this correlation may also be misleading, since it is always possible that the correlation at one level, e.g., whole words, may reflect storage at a different level, e.g., roots.

One way to define the relationship between the frequency of whole words and their components is to consider complex words as the output of a word formation process. In this paper, we restrict our analysis to inflected words, morphologically complex words which we analyze as generated by the combination of a *base* and a single *suffix*. We use the terms *base* and *suffix* throughout, but no alignment with any particular theoretical perspective on word formation is implied. For our purposes, all theories which recognize a connection between words sharing the same base (e.g., *laughed* and *laughing*) are isomorphic.

Whole word frequency can be expressed as the product of two components:

$$p(\text{word}) = p(\text{base, suffix}) = p(\text{base}) \cdot p(\text{suffix}|\text{base})$$

In prose, the probability of an inflected word is the probability of the joint occurrence of the word's base and inflectional suffix. For example, the probability of *laughed* would be the probability of the base *laugh* multiplied by the probability with which *laugh* takes the suffix *-ed*. From this equation, it is apparent that frequency effects at one level might be explained at another. It might seem, as Baayen *et al.* (2003) write, that whole word frequency effects pose a challenge to full decomposition models, which posit no whole word storage, but this might reflect whole word storage, a correlated measure such as base frequency, or even a combination of the two. The correlation between whole word and base frequency is attenuated in languages like English, which have limited inflectional morphology, and in high word frequency forms, since these are the whole words that account for the most variance in base frequency counts. This is particularly important given high word frequency forms in Dutch and English form the empirical base for the whole word storage hypothesis.

## 3 Method

Our goal is to explain the contradictory findings concerning frequency effects, and more generally to establish a firm empirical base on which to construct cognitive models of lexical access. We suspect that the many divergent results cannot only be attributed to differences in task or language; rather, they reflect the poor generalizability of classic lexical decision experiments. Most of the work cited above relies on experiments with a few dozen target stimuli, often taken from a single morphological class. All too frequently, little attention is paid to variation in reaction time between subjects or items. In this study, we revisit frequency effects in lexical decision using a much larger data set and a more principled statistical methodology.

### 3.1 Materials

To provide the largest sample size possible, we use data from the English Lexicon Project (ELP; Balota *et al.* 2007), a visual lexical decision "megastudy" with over 40,000 unique words and nonwords. Baayen *et al.* (2007), Yap & Balota (2009) model real word reaction times for subsets of the ELP data, but average per-item reaction times. As Baayen (2004) notes, this produces an unnecessary reduction in statistical power (i.e., an increase in Type II error). Consequently, we conduct all our analyses at the trial level.

We model reaction times in a subset of this database filtered by pre-determined exclusion criteria at the level of trial, subject, and item. All trials with incorrect responses are excluded. No trials are excluded or truncated based on reaction time; Ulrich & Miller (1994) advise that the effect of reaction-time exclusions is highly unpredictable and that the cost of excluding "real" data may outweigh any benefits of excluding outliers. We also exclude data from a small number of speakers who reported that English was not their first language at test time. We analyze only items which are real words consisting of a simplex base and a regular inflectional suffix (past tense *-d*, progressive and gerund *-ing*, and noun plural and verbal agreement *-s*) according to the ELP morphological coding. This excludes free bases, compounds, derivatives, as well as irregularly inflected words. Items which do not share a base with any other word in the ELP morphological coding were excluded, as their base frequency could not be accurately estimated. The resulting data comprises 201,856 trials, 6,684 unique words, and 772 subjects.

### 3.2 Predictors

While a vast number of reaction time predictors have been proposed in the literature, here we choose predictors which either reflect well-known properties of speeded reading or measure possible frequency effects at different levels of lexical memory.

**Orthographic word length** Word length has long been recognized to have an inhibitory effect on visual word recognition, as longer words may require additional saccades even for shallow processing. While many measures of word length have been proposed, New *et al.* (2006) find that squared orthographic length (i.e., number of characters squared) is most closely correlated with lexical decision latencies in the ELP data, and we adopt this measure.

**Number of syllables** We also include another length measure, syllable count, as New *et al.* (2006) find that this has a robust inhibitory effect on visual recognition which is partially independent of squared word length.

**Orthographic neighborhood density** Lexical decision latencies are longer when the target (whether word or non-word) is similar to many other existing words. Yarkoni *et al.* (2008) propose a new measure of similarity called Orthographic Levenshtein Distance (OLD20) which is more closely correlated with latencies in the ELP data than other measures proposed in the literature. We include this predictor as reported in the ELP database.

**Word and base frequency** Word and base frequency represent some of the most important predictors of lexical decision latencies and it has recently been recognized that analysis of word and base frequency effects are sensitive to the choice of frequency norms. We use the SUBTLEX-US frequency norms, derived from a a corpus of 51 million tokens of American English movie subtitles. Brysbaert & New (2009) report that these norms are more closely correlated with a number of behavioral measures than other popular frequency norms, including the Brown Corpus (Kučera & Francis 1967), CELEX2 (Baayen *et al.* 1996) and HAL (Burgess & Livesay 1998). For each word, bases were identified according to the ELP database morphological segmentations. Base frequency is computed by summing the frequencies of all the words in the ELP database that share the base. As is standard, frequency measures were log-transformed before they were entered into the model.

Most of the prior studies dichotomize these measures by, for example, grouping together all words with "high base frequency" and ignoring frequency-related variance within that group. It has long been known, however, that that dichotomizing continuous measures like frequency greatly reduces statistical power and increases the rate of Type I error (e.g., Cohen 1983, Baayen 2004). We do not include measures of *subjective* (subject-reported) *frequency*, as Brysbaert & Cortese (2011) find that that subjective frequency ratings are not needed to model lexical decision latencies if high-quality "objective" frequency norms are available.

**Suffix conditional probability** Some inflected forms of a base are more common than others, for example *admired* is roughly five times more frequent than *admires*; equivalently, the base *admire* is roughly five times more likely to take the past tense suffix *-d* than the 3sg. agreement suffix *-s*. *Suffix conditional probability* is defined as the probability that a suffix appears given a particular base:

$$p(\text{suffix}|\text{base}) = \frac{p(\text{base, suffix})}{p(\text{base})}$$

This measure is also log-transformed before modeling.

**Other fixed effects** We incorporate a number of fixed effects unrelated to the variables of interest in this study but known to influence reaction time. *Trial number* controls for any effects of fatigue, and subject *education level* and *gender* account

| | SL | NS | OLD | WF | BF |
|---|---|---|---|---|---|
| Squared length (SL) | – | | | | |
| Number of syllables (NS) | **.779** | – | | | |
| Ortho. Levenshtein Distance (OLD) | **.843** | **.700** | – | | |
| Whole word frequency (WF) | −.096 | −.063 | −.092 | – | |
| Base frequency (BF) | −.134 | −.091 | −.132 | **.600** | – |
| Suffix conditional prob. (SCP) | .059 | .050 | .084 | .074 | −.098 |

**Table 1:** Item-level Pearson correlations between the continuous fixed effects, with non-trivial correlations ($|r| \geq .3$) in bold.

for gross differences between subjects. Identity of the suffix was included as a sum-coded fixed effect.

### 3.3 Modeling procedure

We use mixed effects linear regression with trial log RT as the dependent variable. We use a maximal random effects structure (Barr *et al.* in press). This includes subject-level random intercepts and slopes for each fixed effect. While they are appropriate for this experimental design (Clark 1973), item random intercepts are not included due to convergence failures during modeling. To test the significance of individual predictors, we use the log-likelihood ratio test to compare the full model to a model with the same random effects structure but with the corresponding fixed effects removed.

### 3.4 Residualization

As already discussed, there are non-trivial correlations between base and word frequency; in fact, as Table 1 shows, partial multicollinearity afflicts all item-level predictors except suffix conditional probability. The predictors must be made orthogonal to achieve numerical stability and efficient convergence with the mixed effects model estimation procedure. We accomplish this with a technique known as *residualization*. In the case of two non-trivially correlated predictors $X_i, X_j$, one predictor is selected to enter the regression first, and then a linear model with the latter as the dependent variable and the former as the independent predictor:

$$X_j = \beta X_i + \varepsilon$$

A new predictor $X_j'$ is then defined to be equal to $\varepsilon$, the residual error, and entered into the model. Since the previous equation describes a model of $X_j$ in terms of a linear scaling of $X_i$, this is the part of $X_j$ which is *not* explained by—is uncorrelated with—$X_i$. Ford *et al.* (2010) use this technique to study the effects of base and word frequency on a lexical decision task: they enter word frequency first and residualize base frequency with respect to it. As Gorman (2010) shows, this procedure can be carried on indefinitely for multiple correlated predictors, as follows:

$$\begin{aligned} X_j' &= \text{residual}(X_j|X_i) \\ X_k' &= \text{residual}(X_k|X_i, X_j') \\ X_l' &= \text{residual}(X_l|X_i, X_j', X_k') \end{aligned}$$

This technique is often sensitive to the order in which variables are residualized. In our models, we enter non-frequency predictors (squared orthographic length, number of syllables, OLD20) first, de-emphasizing the overall effect of the frequency components. We then enter the different components of frequency in varying orders, depending on the hypothesis being tested.

## 4 Evaluating whole word frequency as a storage diagnostic

Alegre & Gordon (1999, henceforth AG99), explore frequency effects across a broad range of frequencies and argue for a dual-route model in which "whole-word frequency effects [are] obtained for items in the higher end of the frequency distribution but not for items in the lower end" (*ibid.*:43). In their Experiment 2, they find whole word frequency effects only among higher frequency words and conclude that recognition based on whole word storage is available for regularly inflected words with whole word frequencies above 6 tokens per million words. However, they interpret the lack of whole word frequency effects among lower-frequency words as evidence that they must be decomposed during recognition. In this section we use trials of the same items from the ELP to replicate Experiment 2. Unlike AG99, we find robust frequency effects among low and high frequency words.

### 4.1 Methodological challenges in exploring frequency effects

There are a number of reasons to be concerned that the findings of AG99 may not be replicable: the use of low quality frequency norms, failure to address multicollinearity among predictors in their linear model, and the arbitrary dichotomization of frequency using a median split. Alegre & Gordon use the Kučera & Francis (1967) norms, which would account for significantly less variance in their data (Brysbaert & New 2009) than the then-available CELEX2 (Baayen *et al.* 1996) norms or the newer SUBTLEX-US norms. For the items used in their Experiment 2, the Pearson correlation between the Kučera & Francis and SUBTLEX-US frequency norms is 0.604, surprisingly low for two measures of whole word frequency. Additionally, they perform multiple regression with orthographic length, base frequency—called *cluster frequency* in their study—and whole word frequency in the same model without any controls for multicollinearity. Analyzing trials of a subset of their items from their Experiment 2 using ELP trials and their modeling strategy results in a condition number ($\kappa$) of 132; $\kappa > 30$ is thought to indicate a problematic degree of multicollinearity (Belsey *et al.* 1980).

### 4.2 Effects of frequency

The scale of the ELP data allows us to not merely compare the size of frequency effects across a dichotomized measure of word frequencies, as is usually done, but also to examine the effect of frequency across a large range of word frequencies using local regression (LOESS). To obtain a more accurate picture of the average effect of frequency across items, whole word frequency is plotted against the per-item averages of residuals of a baseline model in Figure 1. The residuals are computed by modeling subject-level variance in lexical decision trials using fixed effects for number of trial, education, and gender, and a per-subject random effect. The mean

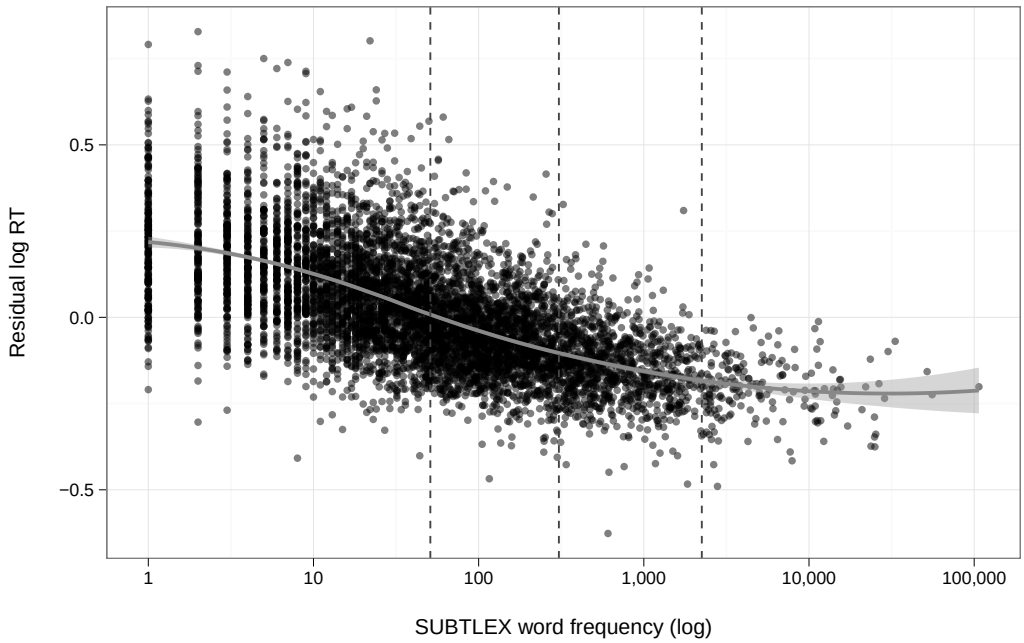


**Figure 1:** LOESS curve fit of whole word frequency against mean per-item residualized reaction times.

residual is computed for each item. Then, log whole word frequency is fit with a LOESS curve. The dashed lines delimit the regions of interest in AG99's Experiment 2, regions equivalent in frequency to the 1–6 and 7–44 per million frequency ranges in the Kučera & Francis norms.

Visual inspection of the frequency ranges of interest in Figure 1 reveal an immediate incongruity between the pattern observed in our data set and AG99; in our data, the magnitude of the frequency effect appears to in fact be *larger* in the lower frequency range. Modeling confirms that for ELP items, there is a significant effect of whole word frequency in the low range ($\chi^2_{LR}(1) = 147.73$, $p = 5.45 \times 10^{-34}$). AG99 find no significant whole word frequency effect in the lower range; modeling all items used in AG99 for which there are trials in the ELP using the Kučera & Francis norms and the model structure described in §3 leads to significant effects in both ranges (low, $\chi^2_{LR}(1) = 4.05$, $p = .0441$; high, $\chi^2_{LR}(1) = 4.04$, $p = .0445$). Thus, contrary to AG99's findings, there is indeed a significant effect of whole word frequency among low frequency items, even when using a subset of the same items used in their study. Surprisingly, lower frequency items participate *more strongly* in this effect than higher frequency ones. This raises doubts as to whether there exists any point below which whole word frequency effects are not observed. As shown in Figure 1, lowering the frequency range of interest does not decrease the size of whole word frequency effects; the only way to select a frequency range in which no frequency effects would be found would be to select the highest frequency items. Applying the logic of AG99, this would lead to the nonsensical conclusion that the highest frequency items are computed online while the rarest words are stored.

A crucial question to consider is why we find such different results in our study

than in AG99. While we cannot identify a single change in methodology between our study and theirs responsible for the difference, the larger sample size and appropriate handling of multicollinearity imply a lower Type II error rate for our study, allowing us to detect effects missed in Alegre and Gordon's study. Even across AG99's experiments, poor statistical power is apparent; among low frequency items they find an effect of base frequency in Experiment 4 but not in Experiment 2. In summary, high-quality frequency estimates and more principled statistical methodology leads us to conclude that there is no evidence that whole word frequency effects are qualitatively different for high frequency and low frequency words.

### 4.3 Modeling the effects of decomposition

Dual-route models (e.g., Baayen *et al.* 1997) predict that the nature of frequency effects observed should depend on the route used for lexical access. For items that trigger the whole word pathway, whole word frequency effects should be observed. For items that trigger the decompositional route, frequency effects should be observed on the decomposed components, bases and suffixes. The challenge in formalizing such a model and testing this claim is to identify the criteria for determining which of the two pathways is used, and when. The most important potential criterion is thought to be whole word frequency, but as we have demonstrated, there is no evidence for a behavioral distinction between high and low frequency items. No other consistent and testable criterion has emerged, making the claims of a dual-route model unverifiable.

We are able to test whether we see effects of decomposition across the broad frequency range recorded in the ELP data. This can be done by comparing a model with only whole word frequency, a non-decompositional model, against one with base frequency and suffix conditional probability, a decompositional model. If we find that the model with the decompositional predictors provides a better fit, this demonstrates that decomposition on the whole provides a better account for the data. This in itself would not preclude a dual-route model, but it would show that the effect of base frequency is not "marginal" as Baayen *et al.* (2007) claim.

Though it is not obvious, the simpler model, with a whole-word frequency predictor, is nested by a model in which whole word frequency is represented by the independent predictors of base frequency and suffix conditional probability. This is because mixed effects estimation is not required to assign them different coefficients, and the decomposition of whole word frequency can be re-written as:

$$\beta_i \, log_2 \, p(\text{word}) = \beta_i \, log_2 \, p(\text{base}) + \beta_i \, log_2 \, p(\text{suffix}|\text{base})$$

In other words, log word frequency times some constant $\beta_i$ is the sum of $\beta_i$ times log base frequency and $\beta_i$ times log suffix conditional probability. Comparison of maximal mixed effects models for the whole word and decompositional models results in a significantly better fit for the decompositional model ($\chi^2_{LR}(7) = 411.74$, $p = 7.24 \times 10^{-85}$),[2] showing that the decompositional model provides a better ac-

[2]The apparently large number (7) of degrees of freedom in the model comparison stems from the use of correlated random effects in a maximal model. One degree of freedom is required for the additional fixed effect, one for the additional per-subject random slope, and five for the correlation of the random slope with other random effects.

| | Estimate | Std. err. | $p$-value |
|---|---|---|---|
| Intercept | 9.28 | $1.45\times10^{-2}$ | (n.a.) |
| Trial number | $-2.76\times10^{-5}$ | $1.01\times10^{-6}$ | $2.40\times10^{-164}$ |
| Education | $4.81\times10^{-3}$ | $1.24\times10^{-3}$ | $1.24\times10^{-4}$ |
| Gender = Male | $-8.93\times10^{-2}$ | $1.01\times10^{-2}$ | $1.90\times10^{-18}$ |
| Suffix = *-ed* | $9.51\times10^{-3}$ | $1.50\times10^{-3}$ | $3.28\times10^{-44}$ |
| Suffix = *-ing* | $1.22\times10^{-2}$ | $1.81\times10^{-3}$ | (n.a.) |
| Suffix = *-s* | $-2.18\times10^{-2}$ | (n.a.) | (n.a.) |
| Squared length | $2.99\times10^{-3}$ | $7.08\times10^{-5}$ | $2.36\times10^{-208}$ |
| Ortho. Levenshtein Distance | $8.70\times10^{-2}$ | $3.57\times10^{-3}$ | $2.04\times10^{-101}$ |
| Resid. number of syllables | $2.75\times10^{-2}$ | $2.30\times10^{-3}$ | $2.00\times10^{-30}$ |
| Resid. base frequency | $-2.98\times10^{-2}$ | $5.09\times10^{-4}$ | $1.15\times10^{-282}$ |
| Resid. suffix conditional prob. | $-2.66\times10^{-2}$ | $6.54\times10^{-4}$ | $1.33\times10^{-190}$ |

Model: $log_2$`(RT) ~ Trial Number + Gender + Education + Suffix + Squared Length + Resid. OLD + Resid. Num. Syllables + Resid.` $log_2$ `Base Frequency + Resid.` $log_2$ `Suffix Conditional Probability + (Squared Length + Resid. OLD + Resid. Num. Syllables + Resid.` $log_2$ `Base Frequency + Resid.` $log_2$ `Suffix Conditional Probability | Subject)`

**Table 2:** Coefficients, associated statistics, and the model specification for the decompositional model.

count of the ELP data: the effect of decomposition is certainly not "marginal".

Table 2 summarizes the decompositional model; following Barr *et al.* (in press), $p$-values were computed via log-likelihood ratio test comparing the full model and a model with the fixed effect of interest removed, but random effects structure held constant.[3] Modeling log RT produces in a better fit than modeling untransformed RT because of the positive skewed nature of RTs. However, a log response makes coefficient interpretation difficult. By using the same model but with an untransformed response, it is clear that the effects of interest are of acceptable size: doubling base frequency decreases RT by 15.9ms, doubling suffix conditional probability decreases it by 14.7ms.

## 4.4 Discussion

The fact that AG99's failure to find an effect of whole word frequency in low frequency ranges was so widely cited as evidence for a dual-route model merits further discussion. Their evidence was simply a positive result in one set of items and a negative result in another set. However, if we assume for the sake of argument that decomposition always occurs during lexical access, we would expect a number of studies to show the same result due to the risk of Type II error; some studies will always fail to find the true effect, especially when plagued with methodological problems (e.g., failure to control for multicollinearity) inflating the Type II error rate. Thus, studies which rely on negative results and which are limited to a single

[3] In Table 2, we provide the $p$-value associated with testing for all levels of the sum-coded suffix predictor and the std. error for the two levels estimated by the model.

set of items and a single experiment (e.g., AG99, Baayen *et al.* 1997) do not provide evidence for the dual-route hypothesis; they are necessary but not sufficient.

In order to have strong evidence for a dual-route model, we would need to satisfy a number of requirements. First, there would need to be a testable set of criteria for determining which items are stored and which are composed. Second, it would need to be demonstrated on a data set such as the ELP that when those criteria are applied, the resulting model fits the data better than a decompositional model.

We are not able check for evidence of a dual-route model because we do not believe there are sufficient experimental results informing us about the conditions under which whole complex words are stored. For example, Baayen *et al.* (1997) give evidence that whole word storage likely plays a part in the handling of certain high frequency items and a specific inflection of a specific syntactic category in Dutch based on potential ambiguity of a specific suffix, but we do not yet have evidence of a generalized motivation for when whole word storage of morphologically complex words occurs. We do not present a model that excludes whole word storage; whole word storage can be incorporated if it has sufficient motivation. However, as noted above, this would require an explicit model of what is stored, something that is nothing more a promissory note after decades of the dual-route hypothesis.

In its current form, the dual-route hypothesis is unfalsifiable. There are no generalizable, explicit models of which forms are stored or computed; when decision criteria have been specified, they are either demonstrably incorrect (e.g., Alegre & Gordon 1999), or ad hoc, without any test of generalization beyond the particular items used in an experiment (e.g., Baayen *et al.* 1997, Betram *et al.* 2000).

## 5 General discussion

### 5.1 Other word classes

This study has focused on regularly inflected words. However, these results bear on the processing of other types of complex words, insofar as these behave like regularly inflected words in lexical decision tasks.

As vividly demonstrated by the famous *wug*-test (Berko 1958), speakers extend regular patterns to novel roots but only very rarely do so with irregular patterns. In contrast, the morphophonological idiosyncrasies of irregularly inflected words make them a likely candidate for storage. Kelliher & Henderson (1990) find that irregulars exhibit base frequency effects when whole word frequency is held constant. Some studies find that irregular primes produce as much facilitation as regular primes (e.g., Allen & Badecker 2002, Fowler *et al.* 1985, Meunier & Marslen-Wilson 2004, Orsolini & Marslen-Wilson 1997, Stockall & Marantz 2006), whereas others report that irregulars prime less than regulars (e.g., Feldman & Fowler 1987, Kempley & Morton 1982, Marslen-Wilson *et al.* 1993, Napps 1989, Stanners *et al.* 1979a, Sonnenstuhl *et al.* 1999).

Similarly, derived words may have morphophonological, syntactic, or semantic idiosyncrasies suggestive of whole word storage. However, Colé *et al.* (1989, 1997), Ford *et al.* (2010), Taft (1979), and Taft & Ardasinski (2006) find the same independent effects of word and base frequency effects for derivationally-related words also found in inflection. Further, many studies report facilitation with deriva-

tional primes (e.g., Emmorey 1989, Forster & Azuma 2000, Henderson *et al.* 1984, Marslen-Wilson *et al.* 1994, Marslen-Wilson *et al.* 2008, Rastle *et al.* 2000, Stanners *et al.* 1979b, Taft & Kougious 2004). Raveh & Rueckl (2000) report that inflectional and derivational priming effects are of approximately the same magnitude.

While there is not yet a consensus in this literature, we conclude that there is not yet sufficient evidence to reject the null hypothesis that derived or irregularly inflected words are processed in the same fashion as regularly inflected words.

### 5.2 Whole word storage in phonology

Our demonstration that the correlation between whole word frequency and processing time neither entails whole word storage nor imperils decomposition-based recognition lessens the appeal of dual-route theories to phonologists, who have recently adopted the hypothesis of frequent whole word storage to derive correlations between word frequency and variable phonological processes (e.g., Bybee 2001) and lexically-specific phonological alternations (e.g., Hayes & Londe 2006).

## 6 Conclusions

This set of experiments presents the first large-scale validation of decompositional models of morphological processing. We have demonstrated that the most widely cited evidence for the absence of frequency effects in low-frequency regulars (Alegre & Gordon 1999) cannot be replicated and was most likely the result of a Type II error from poor frequency estimates and statistical practices. We have shown that the correlation between whole word frequency and processing time is not inconsistent with full decompositional models of lexical access and that a decompositional model provides a good account for a large set of lexical decision data. While we do not rule out a dual-route model, we conclude that the dual-route literature has thus far failed to produce a testable hypothesis that could be used to compare dual-route and decompositional models more generally. Even the impoverished morphology of English can provide some preliminary steps towards the proper understanding of morphological complexity. The recent availability of "lexicon projects" in many other languages opens further avenues for research.

## References

Alegre, M., & P. Gordon. 1999. Frequency effects and the representational status of regular inflections. *Journal of Memory and Language* 40.41–61.

Allen, M., & W. Badecker. 2002. Inflectional regularity: Probing the nature of lexical representation in a cross-modal priming task. *Journal of Memory and Language* 46.705–722.

Baayen, R.H. 2004. Statistics in psycholinguistics: A critique of some current gold standards. *Mental Lexicon Working Papers* 1.1–45.

——, T. Dijkstra, & R. Schreuder. 1997. Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language* 37.94–117.

——, J.M. McQueen, T. Dijkstra, & R. Schreuder. 2003. Frequency effects in regular inflectional morphology: Revisiting Dutch plurals. In *Morphological structure in language processing*, ed. by R.H. Baayen & R. Schreuder, 355–390. Berlin: Mouton.
——, R. Piepenbrock, & L. Gulikers, 1996. CELEX2. LDC96L14: Linguistic Data Consortium, Philadelphia.
——, & R. Schreuder. 1999. War and peace: Morphemes and full forms in a noninteractive activation parallel dual-route model. *Brain and Language* 68.27–32.
——, L.H. Wurm, & J. Aycock. 2007. Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. *Mental Lexicon* 2.419–463.
Balota, D.A., M.J. Yap, M.J. Cortese, K.A. Hutchison, B. Kessler, B. Loftis, J.H. Neely, D.L. Nelson, G.B. Simpson, & R. Treiman. 2007. The English Lexicon Project. *Behavior Research Methods* 39.445–459.
Barr, D.J., R. Levy, C. Scheepers, & H.J. Tily, in press. Random effects structure for confirmatory hypothesis testing: Keep it maximal. To appear in *Journal of Memory and Language*.
Belsey, D.A., E. Kuh, & R.E. Welsch. 1980. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons.
Berko, J. 1958. The child's learning of English morphology. *Word* 14.150–177.
Betram, R., R.H. Baayen, & R. Schreuder. 2000. Effects of family size for complex words. *Journal of Memory and Language* 42.390–405.
Brysbaert, M., & M.J. Cortese. 2011. Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *Quarterly Journal of Experimental Psychology* 64.545–559.
——, & B. New. 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* 41.977.
Burani, C., D. Salmaso, & A. Caramazza. 1984. Morphological structure and lexical access. *Visible Language* 18.348–358.
Burgess, C., & K. Livesay. 1998. The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kučera and Francis. *Behavior Research Methods Instruments and Computers* 30.272–277.
Butterworth, B. 1983. Lexical representation. In *Language production II: Development, writing, and other language processes*, ed. by B. Butterworth, 257–294. London: Academic Press.
Bybee, J. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.
Caramazza, A., A. Laudanna, & C. Romani. 1988. Lexical access and inflectional morphology. *Cognition* 28.297–332.
Chan, E. 2008. *Structures and distributions in morphology learning*. University of Pennsylvania dissertation.
Clark, H.H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* 12.335–359.
Cohen, J. 1983. The cost of dichotomization. *Applied Psychological Measurement* 7.249–253.
Colé, P., C. Beauvillain, & J. Segui. 1989. On the representation and processing of prefixed and suffixed derived words: A differential frequency effect. *Journal of Memory and Language* 28.1–13.
——, J. Segui, & M. Taft. 1997. Words and morphemes as units for lexical access. *Journal of Memory and Language* 37.312–330.
Drews, E., & P. Zwitserlood. 1995. Morphological and orthographic similarity in visual word recognition. *Journal of Experimental Psychology: Human Perception and*

*Performance* 21.1098–1116.
Emmorey, K. 1989. Auditory morphological priming in the lexicon. *Language and Cognitive Processes* 4.73–92.
Feldman, L.B., & C.A. Fowler. 1987. The inflected noun system of Serbo-Croatian: Lexical representation of morphological structure. *Memory and Cognition* 15.1–12.
Ford, M.A., M.H. Davis, & W.D. Marslen-Wilson. 2010. Derivational morphology and base morpheme frequency. *Journal of Memory and Language* 63.117–130.
Forster, K.I., & T. Azuma. 2000. Masked priming for prefixed words with bound stems: Does *submit* prime *permit*? *Language and Cognitive Processes* 15.539–561.
Fowler, C.A., S.E. Napps, & L. Feldman. 1985. Relations among regular and irregular morphologically related words as revealed by repetition priming. *Memory and Cognition* 13.241–255.
Goldberg, A.E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
Gorman, K. 2010. The consequences of multicollinearity among socioeconomic predictors of negative concord in Philadelphia. *University of Pennsylvania Working Papers in Linguistics* 16.2.66–75.
Grainger, J., P. Colé, & J. Segui. 1991. Masked morphological priming in visual word recognition. *Journal of Memory and Language* 30.370–378.
Halle, M., & A. Marantz. 1993. Distributed morphology and the pieces of inflection. In *The view from Building 20: Essays in linguistics in honor of Sylvain Bromberger*, ed. by K. Hale & S.J. Keyser, 111–176. Cambridge: MIT Press.
Hankamer, J. 1992. Morphological parsing and the lexicon. In *Lexical representation and process*, ed. by W.D. Marslen-Wilson, 392–408. Cambridge: MIT Press.
Hayes, B., & Zs. Londe. 2006. Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology* 23.59–104.
Henderson, L., J. Wallis, & D. Knight. 1984. Morphemic structure and lexical access. In *Attention and performance X: Control of language processes*, ed. by H. Bouma & D. Bouwhuis, 211–226. Hillsdale, NJ: Erlbaum.
Howes, D.H., & R.L. Solomon. 1951. Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology* 41.401–410.
Jackendoff, R. 1975. Morphological and semantic regularities in the lexicon. *Language* 51.639–671.
Kelliher, S., & L. Henderson. 1990. Morphologically based frequency effects in the recognition of irregularly inflected verbs. *British Journal of Psychology* 81.527–539.
Kempley, S.T., & J. Morton. 1982. The effects of priming with regularly and irregularly related words in auditory word recognition. *British Journal of Psychology* 73.441–454.
Kučera, H., & W.N. Francis. 1967. *Computational analysis of present-day American English*. Providence: Brown University Press.
Marslen-Wilson, W.D., M. Bozic, & B. Randall. 2008. Early decomposition in visual word recognition: Dissociating morphology, form, and meaning. *Language and Cognitive Processes* 23.394–421.
——, M. Hare, & L. Older. 1993. Inflectional morphology and phonological regularity in the English mental lexicon. In *Proceedings of the 15th annual meeting of the Cognitive Science Society*, 693–698. Princeton: Erlbaum.
——, L.K. Tyler, R. Waksler, & L. Older. 1994. Morphology and meaning in the English mental lexicon. *Psychological Review* 101.3–33.
Meunier, F., & W.D. Marslen-Wilson. 2004. Regularity and irregularity in French verbal inflection. *Language and Cognitive Processes* 19.561–580.
Murrell, G.A., & J. Morton. 1974. Word recognition and morphemic structure. *Journal*

*of Experimental Psychology* 102.963–968.
Napps, S. 1989. Morphemic relationships in the lexicon: Are they distinct from semantic and formal relationships? *Memory and Cognition* 17.729–739.
New, B., M. Brysbaert, J. Segui, L. Ferrand, & K. Rastle. 2004. The processing of singular and plural nouns in French and English. *Journal of Memory and Language* 51.568–585.
——, L. Ferrand, C. Pallier, & M. Brysbaert. 2006. Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin and Review* 13.45–52.
Orsolini, M., & W.D. Marslen-Wilson. 1997. Universals in morphological representation: Evidence from Italian. *Language and Cognitive Processes* 12.1–47.
Rastle, K., M.H. Davis, & W.D. Marslen-Wilson. 2000. Morphological and semantic effects in visual word recognition: A time-course study. *Language and Cognitive Processes* 15.507–537.
Raveh, M., & J.G. Rueckl. 2000. Equivalent effects of inflected and derived primes: Long-term morphological priming in fragment completion and lexical decision. *Journal of Memory and Language* 42.103–119.
Schriefers, H., A. Friederici, & P. Graetz. 1992. Inflectional and derivational morphology in the mental lexicon: Symmetries and assymetries in repetition priming. *Quarterly Journal of Experimental Psychology* 44A.373–390.
Sereno, J.A., & A. Jongman. 1997. Processing of English inflectional morphology. *Memory and Cognition* 25.425–437.
Sonnenstuhl, I., S. Eisenbeiss, & H. Clahsen. 1999. Morphological priming in the German mental lexicon. *Cognition* 72.203–236.
Stanners, R., J. Neiser, W. Hernon, & R. Hall. 1979a. Memory representation for morphologically related words. *Journal of Verbal Learning and Verbal Behavior* 18.399–412.
——, J. Neiser, & S. Painton. 1979b. Memory representation for prefixed words. *Journal of Verbal Learning and Verbal Behavior* 18.733–743.
Stockall, L., & A. Marantz. 2006. A single route, full decomposition model of morphological complexity: MEG evidence. *Mental Lexicon* 1.85–123.
Taft, M. 1979. Recognition of affixed words and the word frequency effect. *Memory and Cognition* 7.263–272.
—— 2004. Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology* 57A.745–765.
——, & S. Ardasinski. 2006. Obligatory decomposition in reading prefixed words. *Mental Lexicon* 1.183–199.
——, & K.I. Forster. 1975. Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior* 14.638–647.
——, & K.I. Forster. 1976. Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior* 15.607–620.
——, G. Hambly, & S. Kinoshita. 1986. Visual and auditory recognition of prefixed words. *Quarterly Journal of Experimental Psychology* 38A.351–366.
——, & P. Kougious. 2004. The processing of morpheme-like units in monomorphemic words. *Brain and Language* 90.9–16.
Ulrich, R., & J. Miller. 1994. Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General* 123.34–80.
Yap, M.J., & D.A. Balota. 2009. Visual word recognition of multisyllabic words. *Journal of Memory and Language* 60.502–529.
Yarkoni, T., D. Balota, & M. Yap. 2008. Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin and Review* 15.971–979.