Jeff Parker[1]

Word Frequency and Processing:
Why the Brain Stores Some Words Whole and Others in Parts

Introduction

Discussion about the structure of the lexicon has primarily focused on morphologically complex words. Theories about the lexicon assume that certain items are stored, such as morphologically simple words, e.g. *hero* and *govern*, and derivational suffixes, e.g. *-ism* and *-ment*. Given these assumptions, the majority of arguments discuss the status of morphologically complex words, e.g. *heroism* and *government*. Theories posit different levels of parsing and storage. The extent to which theories accept parsing as a active process during lexical access ranges from classical approaches which assume all morphologically complex words are parsed, to theories which suggest all words, simple and complex, are stored whole.

In this paper I argue that we must consider both simple and complex words as candidates for parsing during lexical access, a concept not previously discussed. I further argue that the base for processing is rooted in the relative frequency between base words and their complex counterparts rather that the traditional view of morphological complexity. To illustrate this I first present a concise review of previous literature including addressing an area of lexical access not previously discussed, namely the status of simple words which are less frequent than their complex counterparts (Section 1). I provide three possible analyses for this type of word (Section 2) and then present my methodology (Section 3) and data (Section 4). I finish with my conclusions and the further implications of this work (Section 5).

1 Previous literature

The structure of the lexicon is a much debated topic in morphology. Much of the debate

[1] This paper is a portion of a larger project done in collaboration with Lauren Ressue, Robert Reynolds and Michael Phelan. A special thanks to Professor Andrea D. Sims for her support, comments and encouragement.

centers around what forms are stored in the lexicon in opposition to forms that are, in some way or another, parsed. Research about morphologically complex words has generally focused on two opposing access methods: parsing and direct access of stored items. Traditionally it is assumed that morphologically simple words are always stored in the lexicon. Thus, words that cannot be broken into parts, like *govern* and *hero*, maintain their own entries in the lexicon. Likewise derivational affixes, like *-ment* and *-ity*, are assumed to be stored in the lexicon. Given these assumptions there are multiple possible analysis for complex words such as *government* and *heroism*. One possibility is that because the component parts of these words are stored independently, that the complex words are processed as a combination of two parts, e.g. *hero+ism*. In this traditional analysis *hero* acts as the base for processing *heroism* and its meaning is computed from the meaning of the parts. Other analyses have gone to an opposite extreme in suggesting that all words, both simple and complex words, are stored in the lexicon with possibly interconnecting relations. This theory is called the Full-listing Hypothesis (Butterworth 1983). The Full-listing hypothesis has found criticized on a number of grounds. Hankamer (1989) showed that highly agglutinative languages, such as Turkish, would demand unreasonable amounts of storage - nine million possible permutations for every noun. Another criticism of a full-listing based approach is that it does not account for various priming effects found in lexical decision tasks. If all word forms were stored independently, then derivational factors should not affect access times in priming experiment; however, many studies attest to significant differences between access times for morphologically simple and complex words (e.g. Marslen-Wilson et al., 1994; Schreuder & Baayen, 1997). The Full Listing Hypothesis does not explain this systematic variation.

Taft and Forster (1975) produced a seminal study in which they argued that in a lexical

decision task, prefixed words are analyzed into their constituent morphemes before lexical access occurs. A later study added evidence indicating that in lexical access, parsing is more likely with suffixes than prefixes (Segui & Zubizarreta, 1985). This study addresses relationships between words derived by suffixation.

Lexical decision times are affected by a number of factors; namely semantic transparency (Marslen-Wilson et al. 1994), orthography (Chateau et al. 2001), relative frequency (Gurel 1999, Hay 2001), semantic relation (Raveh 2002), allomorphy (Jarvikivi et al. 2006), productivity (Bertram et al. 2000), and prosody (Kemps et al. 2005). Thus, a complex interaction of factors seems to determine the likelihood of a word being accessed directly or being parsed. For this study I focus on the affects of one factor: frequency. Using our example above, we can see the frequency of simple words relative to the frequency of the complex words differs across words pairs, see Table 1.

Table 1. Frequency distribution of two words pairs[2]

Simple word (frequency, instances per million)	Complex word (frequency, instances per million)
govern (5.78)	government (61.99)
hero (22.46)	heroism (1.19)

Hay (2001) illustrates that not all complex words are equal in their likelihood of being parsed based on the difference in relative frequency between the simple word and the complex word. For example, while both *heroism* and *government* are complex words in the traditional sense (i.e. can be broken into parts), they do not have equally frequent bases. Words that are frequently accessed maintain a higher resting activation level than words that are accessed only rarely. Thus, given the difference in frequency between *hero* and *govern*, we should not expect them to be equally active as bases for processing. Hay shows that when a morphologically

[2] Frequency counts for English words are from the British National Corpus (100 million words). See http://www.natcorp.ox.ac.uk/ for details. Accessed March 2011.

complex word is more frequent than its morphologically simple base, as is the case with *government*, the complex form will bias towards storage, i.e. being accessed as a whole, rather than be accessed through its relatively infrequent base. Given this, we expect *hero* to act as a base for processing *heroism*, but do not expect *govern* to act as a base for *government*. The likelihood that a morphologically complex word will be parsed is dependent on how frequent it is in comparison to its base.

2 Three possible access paths for *govern*

If *government* is stored independent of its base, we should ask an additional question: what happens to its less frequent base *govern*? Three different analysis can be supported in response to this question; each will be dealt with in turn.

One possibility is that *govern* serves as the base for *government.* This analysis corresponds with the traditional notion that simple words are stored and complex words are parsed. However, this outcome contradicts Hay's work that suggests that relatively more frequent complex words bias towards storage. While I do not expect this to be the correct outcome I include it because it provides a logical possibility and corresponds to the traditional view of morphological storage and parsing, namely that simple words are stored and complex words are parsed.

A second possibility is that both *govern* and *government* are both stored whole. This analysis is compatible with both a full listing hypothesis and Hay's results. However, it is important to note that while this prediction is compatible with Hay's description of frequency it is not a necessary outcome of her conclusion. Hay's predicts that *government* will be stored independent of *govern* but the difference in frequency says nothing about how *govern* will be accessed when *government* becomes independent. Thus, while this prediction is possible, it is

prone to many of the downfalls of the Full-Listing Hypothesis. I include it as a logical possibility but not as my prediction.

A third possibility that has not been suggested previously in the literature is that *government* acts as a base for *govern*. Given that more frequent words have a higher level of resting activation we might expect the more frequent word in a pair to act as the base despite its morphological complexity in the traditional view. This prediction is a logical extension to Hay's claim that more frequent words maintain their own lexical entry; however, it adds the assumption that all less frequent words are dependent on other lexical items. This prediction implies that the base for processing is rooted in relative frequency between similar forms rather than on traditional complexity. This suggests that the relatively less frequent items, e.g. *heroism* and *govern*, are both accessed via parsing despite the fact that one is simple and one is complex. This prediction suggests that *heroism* is accessed as *hero* + *ism* and *govern* is accessed as *government* - *ment*, a subtractive morphological process.

Though it has never been suggested for processing, the concept of subtractive morphology has been (controversially) proposed for inflectional morphology. Haspelmath (2002) proposes the French adjective *blanche* 'white-FEM' as the base for *blanc* 'white-MASC'. Additionally we know that similar processes occur historically. For example, *commune* is known to have been created due to an (incorrect) reanalysis of *community* as 'commune + ity' which is historically inaccurate. Speakers created a new word, *commune*, by recognizing a 'base' (even though it was not such historically) in *community*. Thus, we might expect a similar process to occur during processing. Speakers could access a word from a more complex but related form by stripping off the suffix during access, i.e. *govern* could be accessed as *government* - *ment*.

Given these three possible analysis for *govern*, I now focus on what priming effects we

expect from each of these analysis in a lexical decision task. In each case we expect a priming effect when the prime is acting as the base for the target word. From previous studies we know that a simple word, e.g. *hero*, which is more frequent than its complex counterpart, e.g. *heroism*, will act as the base during lexical access. Thus, as we see in Table 2 (below) we expect a priming effect when *hero* is the prime for *heroism*[3]. For *government/govern* we have different expectations for priming in each prediction. In Analysis 1, *govern* acts as the base and therefore primes *government*. In Analysis 2 both words are stored independently and no priming takes place. In Analysis 3, *government* acts as the base and therefore primes *govern*. Here we can see that the predictor for the base in Analysis 1 is the traditional view of complexity but the predictor of the base in Analysis 3 is frequency.

Table 2. Priming expectations for possible analyses of the relationship between *govern* and *government*

Prime	Target	Analysis 1	Analysis 2	Analysis 3
$\text{Simple}_{\text{High}}$ *hero*	$\text{Complex}_{\text{Low}}$ *heroism*	✔	✔	✔
$\text{Complex}_{\text{Low}}$ *heroism*	$\text{Simple}_{\text{High}}$ *hero*	--	--	--
$\text{Complex}_{\text{High}}$ *government*	$\text{Simple}_{\text{Low}}$ *govern*	--	--	✔
$\text{Simple}_{\text{Low}}$ *govern*	$\text{Complex}_{\text{High}}$ *government*	✔	--	--

3 Methodology

To test which of the three analyses is best supported, we constructed a masked priming lexical decision task. Lexical decision tasks have been used widely to investigate lexical access and masked priming has been shown to be most useful for tasks that address priming of morphologically related forms (Forster and Kenneth 1999). In these tasks, we record how quickly speakers access words. In general, words which are stored directly are accessed faster

[3] We might also expect a smaller priming effect in the opposite direction based on the connection between the two words; however, it will be less significant if it occurs at all.

than words that are parsed. Additionally, accessing a complex word is faster when the base word has been access immediately beforehand. We expected a word like *heroism* to be accessed faster if *hero* directly precedes it, even if *hero* is not consciously recognized. Whether or not *govern* is accessed faster when it is preceded by *government* will help us determine which of the 3 analyses discussed above is best supported.

3.1 Stimuli

Our target stimuli consist of two sets of target words, each with 60 word pairs. In the first set of words, the simple words are more frequent than their complex counterparts. For the second set the reverse is true[4]. In addition to the target words, 100 non-words act as fillers. Of these fillers, half are phonologically absurd while the other half are similar to existing Russian words, differing only by one or two graphemes. Examples are presented below in Table 3.

Table 3. Example stimuli

Word Type (example)	Number of stimuli	Examples
$Simple_{High}$ (hero)	30	*tolstyi* 'heavy' (106 ipm)
$Complex_{Low}$ (heroism)	30	*tolstjak* 'heavy person' (4 ipm)
$Simple_{Low}$ (govern)	30	*zavisimyi* 'dependent' (3 ipm)
$Complex_{High}$ (government)	30	*zavisimost'* 'dependence' (49 ipm)
Phonologically possible non-words	50	marakteristik (0 ipm) otošenie (0 ipm)
Phonologically absurd non-words	50	tsotso (0 ipm) dlviaar' (0 ipm)

As seen in Table 4 (below), the stimuli were divided into four lists, one for each group of participants. To ensure no additional priming effects took place, no word pair should be viewed more than once within a given list. Thus, group A was presented with half of the target stimuli, unprimed, and group B was presented with the other half, also unprimed. Recording unprimed reaction times is necessary to establish that any priming effect exists in other groups. Groups C

[4] All frequency counts are from the Russian National Corpus (150 million words), see http://www.ruscorpora.ru/en/index.html. Accessed September 2010.

and D are constructed to test priming effects in both directions (that is complex → simple and vice versa), so that participants in group C saw half the words in each priming condition and group D saw the other half. In total this allows group C and D to make decisions on 30 stimuli for each of the four priming conditions. Each list also included an equal amount of both types of filler words. Filler words are necessary to ensure participants must decide whether the stimuli are real words. In total, no single speaker see the same word twice during the experiment.

Table 4. Contents of Lists

Target Words (# of)	Group A (Unprimed)	Group B (Unprimed)	Group C	Group D
$Simple_{High}$ (60)	30_x	30_y	30_x	30_y
$Complex_{Low}$ (60)	30_x	30_y	30_x	30_y
$Simple_{Low}$ (60)	30_x	30_y	30_x	30_y
$Complex_{High}$ (60)	30_x	30_y	30_x	30_y
Fillers (Phon. possible)	50	50	50	50
Fillers (Phon. absurd)	50	50	50	50
Total	160	160	220	220

To avoid extraneous effects from absolute frequency within the lists, each subset was regularized according to absolute frequency. This was done by assuring that the mean and standard deviation of the frequencies of each section were similar. Lists were also organized, to the extent possible, to minimize differences in the mean word length. See Table 5 below.

Table 5. Regularization of absolute frequency across subsets of stimuli

List	Sublist	Freq. of More Common Form (Standard Deviation)	Freq. of Less Common Form (Standard Deviation)	Relative Frequency	Length in Letters
Simple High Complex Low *ex. geroj - geroizm*	Full	16174.88 (σ = 11200.52)	3507.79 (σ = 3606.05)	0.24	5.98
	Sublist A	15781.65 (σ = 11277.03)	3546.80 (σ = 3883.89)	0.23	6.20
	Sublist B	16568.11 (σ = 11124.02)	3468.75 (σ = 3328.2)	0.25	5.75
Simple Low Complex High *ex. gosudar' - gosudarstvo*	Full	18922.90 (σ = 16163.68)	6348.98 (σ = 8398.17)	0.36	6.78
	Sublist A	18140.85 (σ = 14572.14)	6397.05 (σ = 10289.14)	0.34	6.75
	Sublist B	19704.95 (σ = 17755.22)	6300.90 (σ = 6507.20)	0.37	6.80

3.2 Participants

The participants consisted of 17 native speakers of Russian. The speakers were divided into four groups; each group was given one of the lists described (see Table 4 above).

3.3 Procedure

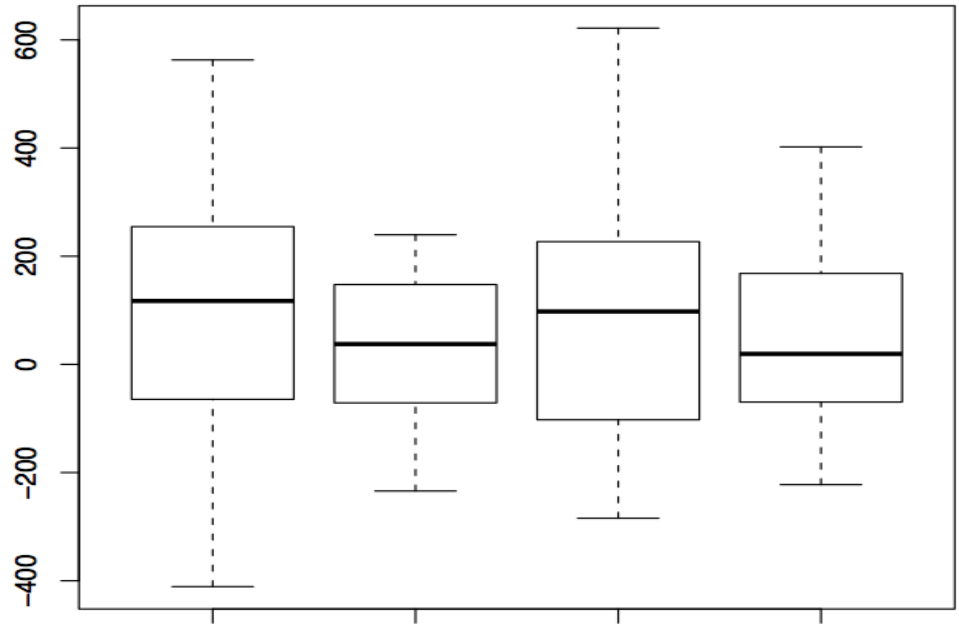
Within the lexical decision task, participants were asked to make judgments on whether a word presented visually on a computer screen was a real word or not, and their response times were recorded using the computer program E-Prime. Participants were shown the tokens and asked to press a labeled button if it was a real word, or press a different one if it was not. The words/non-words were printed in the Russian Cyrillic alphabet. The participants with unprimed lists saw a fixation (+++) for 1 second, followed by the target word for 3 seconds. If the participant had not made a decision in 3 seconds, the next fixation appeared on the screen. Participants who were given primed words saw the same fixation for 1 second, followed by the prime for 30 milliseconds. Even though the primes were only shown briefly, a visual effect was

consciously perceptible[5]. Thus, to maintain a uniform representation of all stimuli within the task, filler words were primed with unrelated filler words. After the prime, the target appeared on the screen for 3 seconds. The stimuli were broken into two blocks, each consisting of 110 tokens. After the first block, participants were given a chance to take a break.

4 Results

Priming effects for each group are shown in Figure A. As one can see, two groups, $Complex_{High}$ priming $Simple_{Low}$ and $Simple_{High}$ priming $Complex_{Low}$, had a greater priming effect than the other two groups. The fact that only high frequency words primed low frequency words is telling. This suggests that the traditional notion of complexity is not an accurate predictor of which word, in a given pair, acts as the base for processing. On the other hand, high frequency words consistently primed low frequency words suggesting that frequency is an accurate predictor in determining the base for processing.

Figure A. Priming effects by word type in milliseconds



[5] Even though a slight effect was perceptible, during the debriefing most speakers reported that they did not notice it.

I now return to the three predictions from Section 2 above. We can see that the priming results support Prediction 3 suggesting that *government* acts as a base for *govern* during processing. This suggests that even simple words, e.g. *govern*, can be parsed during lexical access. Additionally this suggests that complex words not only bias towards storage, as Hay suggests, but they actually acquire the role of the base word. Thus, *government* acts as the base for *govern* during processing.

Table 6. Priming expectations and results

Prime	Target	Prediction 1	Prediction 2	Prediction 3	Results
$\text{Simple}_{\text{High}}$ *hero*	$\text{Complex}_{\text{Low}}$ *heroism*	✔	✔	✔	✔ p = 0.03
$\text{Complex}_{\text{Low}}$ *heroism*	$\text{Simple}_{\text{High}}$ *hero*	--	--	--	-- p = 0.59
$\text{Complex}_{\text{High}}$ *government*	$\text{Simple}_{\text{Low}}$ *govern*	--	--	✔	✔ p = 0.01
$\text{Simple}_{\text{Low}}$ *govern*	$\text{Complex}_{\text{High}}$ *government*	✔	--	--	-- p = 0.37

5 Conclusions

Here I have shown that complex words that are more frequent than their bases, e.g. *government*, are indeed stored independent of their simple counterparts. Moreover, I have shown that such words act as a base for their relatively infrequent simple counterpart, e.g. *govern*, during lexical access. This suggests that the traditional notion of complexity is not an accurate predictor of when a word will act as a base for processing. In contrast, the relative frequency between word pairs is a good predictor of which word will act as the base for processing. This suggests that even traditionally simple words can be 'parsed' via a process of subtractive morphology. More generally these results suggest that frequency is playing a larger role than previously thought in relation to the structure of the lexicon. Words that are less frequent than other morphologically related words, e.g. *govern*, become dependent on more frequent forms during lexical processing.

6 References

Baayen, R. H.; Piepenbrock R.; and Gulikers L. 1995. The CELEX Lexical Database (Release 2). Philadelphia: Linguistic Data Consortium.

Butterworth, B. 1983. Lexical representation. In, Language production: Vol. II. Development, writing and other language processes. ed. by B. Butterworth, London: Academic Press: 257-294.

Forster, Kenneth. 1999. The microgenesis of priming effects in lexical access. Brain and Language 68: 5-15.

Hankamer, J. 1989. Morphological parsing and the lexicon, in Lexical Representation and Process, ed. byWilliam Marslen-Wilson, Cambridge, MA: MIT Press: 392-408.

Hay, Jennifer. 2001. Lexical frequency in morphology: Is everything relative? Linguistics 39(6): 1041-1070.

Hay, Jennifer and I. Plag. 2003. What constrains possible suffix combinations? On the interaction of grammatical processing restrictions in derivational morphology. Natural Language and Linguistic Theory 22: 565-596.

Marslen-Wilson, W.D., L.K. Tyler, R. Waksler, and L. Older. 1994. Morphology and meaning in the English mental lexicon. Psychological Review 101: 3-33.

Meyer, David E. and Roger W. Schvaneveldt. 1971. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. Journal of Experimental Psychology 90(2): 227-234.

Rastle & Brysbaert. 2006. Masked phonological priming effects in English: Are they real? Do they matter? Cognitive Psychology 53: 97-145.

Segui, J., & Zubizarreta, J. 1985. Mental representation of morphologically complex words and lexical access. Linguistics 23: 759-774.