Antoine Tremblay

Empirical evidence for an inflationist lexicon

Abstract: Although Generative and Cognitive grammars both assume that some linguistic forms are stored as wholes while others are strung together from simpler parts, they differ with respect to the units that may be stored in the lexicon, and therefore in the rules that combine them. Within the Generative framework, the determining factor for storage is regularity. In contrast, for Cognitive grammarians frequency of use is an important factor determining whether a form is stored as a whole or (de)composed on-line.

The studies summarized in this paper show that speakers are sensitive not only to the (continuous) frequency/probability of use of single words but also of compositional bigrams (*I really*), trigrams (*I really like*), and quadgrams (*I really like it*). They clearly support the Cognitive view of a redundant, inflationist lexicon that is comprised of a myriad of more or less entrenched linguistic units of different lengths. Implications regarding the entrenchment of N-grams, the opacity of an entrenched N-gram, and the threshold for holistic storage of an N-gram are discussed.

Keywords: the mental lexicon; multi-word sequences; lexical bundles; self-paced reading; immediate free recall; speech production, ERP.

Correspondence address: trea26@gmail.com

1 Introduction

Dual-route models of language assume that some linguistic forms are stored and retrieved as wholes while others are composed from simpler parts via rules. Generative and Cognitive grammars are two examples of dual-route models, which differ, among other things, with respect to what is stored in the lexicon and what is not. Within the Generative framework (e.g., Chomsky 1993; Jackendoff 2002; Halle and Marantz 1993; Marantz 1995), regular forms are computed online (e.g., *ask* + *-ed* → *asked*; *drown* + *-ed* → *drowned*) whereas irregular forms are stored and retrieved as wholes (past tense of *see* is *saw*). However, for Cognitive grammarians, (e.g. Langacker 19987, 1991; Bybee and Hopper 2001; Croft 2007; Goldberg 2009), infrequent forms are generated from smaller parts (e.g., *drown* + *-ed* → *drowned*) whereas frequent ones are stored/retrieved as wholes (past tenses of *ask* and *see* are *asked* and *saw*).

In a similar fashion, regular multi-word sequences such as *I really like it* (i.e. multi-word sequences that can be interpreted on a word-by-word basis, which may or may not have a specialized unitary meaning) are, in the Generative framework, composed through grammatical rules from atomic units but irregular ones such as *grow in the telling* 'the more you tell it, the larger, wilder, better, etc. the story gets' are stored and retrieved as wholes. Under the Cognitive view, frequently used multi-word sequences, whether regular or irregular, are stored and retrieved as wholes whereas infrequent ones are composed via rules from subordinate parts (shorter multi-word sequences, e.g., *I really*, *really like it*, and/or individual words, e.g., *I*, *like*). It follows that the Generative lexicon is a streamlined, atomic one while the Cognitive lexicon incorporates redundancy and is thus inflationist.

Although there is evidence for both the Generative (e.g., Embick et al. 2000; Pinker and Ullman 2002; Grodzinsky and Friederici 2006; Newman at al. 2007; Marslen-Wilson 2007) and the Cognitive views (e.g. Arnon and Snider 2010; Bybee and Scheibman 1999; Bybee and Hopper 2001; Schmitt et al. 2004, Joanisse and Seidenberg 2005; Bybee 2007; Jiang and Nekrasova 2007, Wray, 2002, 2008; Goldberg 2009), the jury is still out on the question of what is stored and what is not (Weinert 2010). In this paper, we present further evidence lending support to the latter view.

2 Lexical bundles and self-paced reading

In Tremblay et al. (2011), we investigated the idea put forth by Biber et al. (1999) that lexical bundles are stored and processed holistically.[1] To do this, we ran three self-paced reading experiments modelled after Schmitt et al. (2004) where we compared four- and five-word lexical bundles such as *I don't know what* and *in the middle of the* to matched non-lexical bundles, for example *I do know what* and *in the front of the*. We reasoned that if lexical bundles are stored and retrieved as single units they would be read more quickly than comparable non-lexical bundles. Indeed, retrieving one lexical bundle from memory ought to be faster than accessing four or five individual words and stringing them together via grammatical rules.

Lexical bundles were defined as having a frequency of occurrence of ten per million or more while five-word bundles had a frequency of at least five per million (Biber et al. 1999). Frequency counts were taken from the *BYU: British National Corpus* (BNC; Davies 2004). Lexical bundles and non-lexical bundles differed in

[1] "Time pressure makes it more difficult for speakers to exploit the full innovative power of grammar and lexicon: instead, they rely heavily on well-worn, prefabricated word sequences, readily accessible from memory". (Biber et al. 1999: 1049. Also see Schmitt 2005).

one word only, which, in lexical bundle sequences, was never shorter or more frequent than in the other type of sequence. This was done in order to ensure that any reading advantages stemming from lexical bundles could not be attributed to these words. Indeed, it is well established that shorter or more frequent words (e.g. *do* and *front*) are processed more quickly than longer or less frequent ones (*don't* and *middle*). It is important to note that in non-lexical bundle sequences the great majority of these words were shorter and more frequent. Thus, if lexical bundles have no psychological reality, non-lexical bundle sequences ought to be read faster. Both types of sequences were embedded in the same carrier sentences (e.g. *I sat _____ bullet train* for the *in the middle of the* / *in the front of the* stimulus pair) in an effort to control for confounding factors that may arise at the sentential level.

Participants either saw a sentence containing a lexical bundle (e.g., *I sat in the middle of the bullet train*) or one containing its non-lexical bundle counterpart (*I sat in the front of the bullet train*). In the first self-paced reading experiment, the first word of a sentence was presented in the centre of a computer screen (see Figure 1A). To see the following word, participants pressed the space bar. Reading times were defined as the time elapsed from the appearance of a word on the screen to the time when a participant pressed the space bar to see the next word. In the second experiment, participants saw the first portion of a sentence. The second portion, which contained the (non-)lexical bundles of interest,

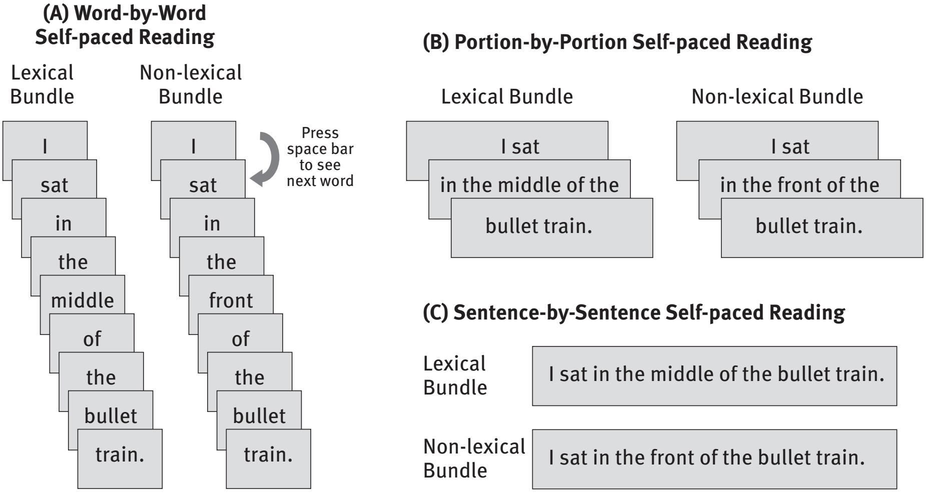

Figure 1: (A) Word-by-word self-paced reading: Participants pressed the space bar to see the next word. (B) Portion-by-portion self-paced reading: Participants pressed the space bar to see the next portion. (C) Sentence-by-sentence self-paced reading: Participants pressed the space bar to see the next sentence.

and the third portion appeared after participants pressed the space bar (once for each portion; see Figure 1B). In the third experiment, a sentence was presented in its entirety and participants pressed the space bar to see the next sentence (see Figure 1C).

In all three experiments, lexical bundles and sentences containing lexical bundles were read faster on average than their non-lexical bundle counterparts (13 milliseconds faster in the first experiment, 124 milliseconds in the second one, and 389 milliseconds faster in the third one; a millisecond is a thousandth of a second) thus lending support to the hypothesis put forth by Biber et al. (1999).

3 Lexical bundles and sentence recall

The lexical bundle facilitatory effect was replicated in two follow-up word and sentence recall experiments reported in Tremblay et al. (2011). In these experiments, which were modelled after Savin and Perchonock (1965), the sentences containing lexical bundle and non-lexical bundle sequences used in the three self-paced reading experiments described above were presented either visually (Figure 2A) or auditorily (Figure 2B). They were followed by a series of six monomorphemic words of equal length and frequency of use (e.g. *date*, *dice*, *male*,

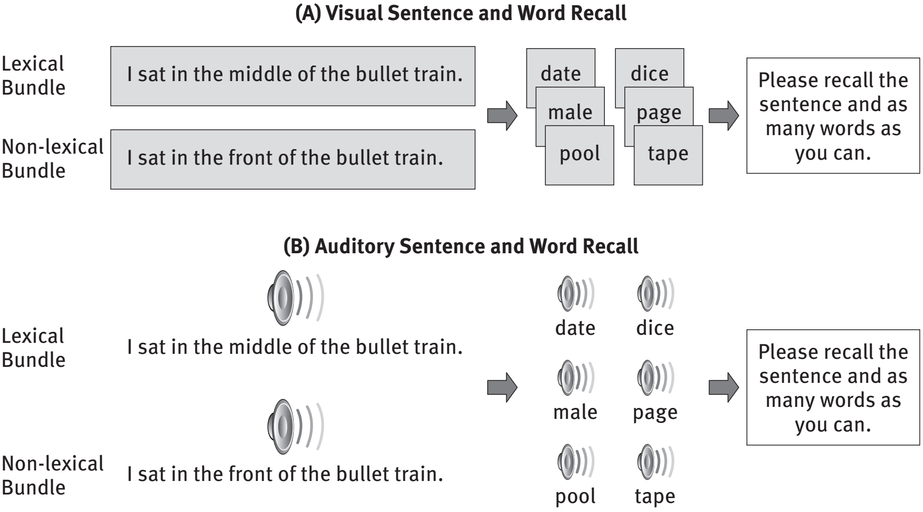
(A) Visual Sentence and Word Recall
Lexical Bundle
I sat in the middle of the bullet train.
Non-lexical Bundle
I sat in the front of the bullet train.
date
dice
male
page
pool
tape
Please recall the sentence and as many words as you can.
(B) Auditory Sentence and Word Recall
Lexical Bundle
I sat in the middle of the bullet train.
Non-lexical Bundle
I sat in the front of the bullet train.
date
dice
male
page
pool
tape
Please recall the sentence and as many words as you can.

Figure 2: (A) Visual sentence and word recall. (B) Auditory sentence and word recall. Participants are asked to remember a sentence and six words for later recall.

page, *pool*, *tape*). The task was to remember the sentence and as many single words as possible.

The idea underlying these experiments is that an individual's working memory can contain up to 7 ± 2 units (Miller 1956). If lexical bundles are stored and retrieved as single units, then committing such a sequence to memory would require one slot. Non-lexical bundles, however, would require four or five slots (one for each word). Therefore, it ought to be easier to correctly recall sentences that contain lexical bundles because one would need to retrieve fewer units from memory. In other words, there would be fewer opportunities for one to retrieve the wrong unit or to substitute a unit that one doesn't remember with another one. Moreover, there ought to be plenty of space left to store additional words when a sentence contains a lexical bundle, but not when it contains a non-lexical bundle.

In both the auditory and visual modalities, sentences containing lexical bundles were more accurately recalled than those that did not. Tremblay et al. (2011) also found that, in the visual modality only, more words were recalled after sentences containing lexical bundles. These results suggest, again, that lexical bundles are psychologically real.

4 From lexical bundles to non-idiomatic multi-word sequences

In the latter five experiments, we strictly looked at the effects of the frequency of occurrence of four- and five-word lexical bundles on processing, keeping other things constant. It is possible, however, that single word frequencies *as well as* the frequency of smaller sequences contained within longer ones also affect processing. This would lend further support to the inflationist lexicon. If th*is is* the case, one may wonder whether linguistic processing occurs in a cascaded, staged fashion (e.g., Levelt 1989, 1992; Jescheniak and Levelt 1994; Levelt et al. 1999) or in parallel (Dell 1986; McClelland 1987; Caramazza 1997; Alario et al. 2002, Hand et al. 2010). In other words, does activation in the brain strictly move from one level of representation to the next (from single words to bigrams to trigrams to quadgrams, e.g. from *I* to *I really* to *I really like* to *I really like it* or from *I really like it* down to the words that comprise it) or does it flow from any level of representation to any other (e.g. the level of activation of *I really* modulates the level of activation of *I really like it* and vice versa; the level of activation of *the* influences the degree to which *in the middle* is activated and vice versa)? Interactions between single word frequencies (*as*), bigram frequencies (*as far*), trigram frequencies (*as* *far as*), and quadgram frequencies (*as far as I*) would indicate that multi-word sequence processing unfolds in parallel, which, one can argue, is in line with the idea of an inflationist lexicon.

Another important point is that probabilistic measures such as mutual information (i.e., the association strength between two or more words, e.g., between *in*, *the*, *way*, and *of*, which is often used in corpus linguistics), and/or probability of use (a.k.a. cloze probability, i.e. the probability of a word occurring given previous context, e.g., the probability of occurring given *in the way*, which is often used for the assessment of native and second language learning), may also affect multi-word sequence processing and may even trump frequency of use (as in, e.g. Ellis and Simpson-Vlach 2009).[2]

4.1 Four-word sequence production

Tremblay and Tucker (2011) investigated these issues in the context of a four-word sequence production task. In this study, we endeavoured to investigate multi-word sequence frequency effects not by comparing high versus low frequency sequences, that is, lexical bundles versus non-lexical bundles, but rather by considering frequency as a continuous variable. To do this, we showed participants four-word sequences with frequencies ranging from very low to very high (in our sample, from 0.3 to 85 occurrences per million words). Whole-sequence (quadgram) frequency counts were taken from the *Contemporary Corpus of American English* (COCA; Davies 2008). Single word, bigram, and trigram frequencies were also obtained. Bigram, trigram, and quadgram mutual information and log probability of use values were computed.

Quadgrams were presented as wholes to participants one at a time (e.g., *at the end of*, *I don't think that*, *in the United States*; see Figure 3). They were asked to say them aloud. We measured the time it took to access linguistic material from the mental lexicon, to plan utterances, and to onset production by measuring the time between the moment a four-word sequence appeared on the screen to the time participants began producing them. We also measured motor program execution times by measuring the time it took participants to produce a sequence.

Both the production onset time analysis and the production duration analysis revealed numerous main effects and interactions between single-word frequencies and the frequency of occurrence, probability of use, and mutual information

[2] Another issue is that there may not be a threshold for holistic storage such as the one proposed, for example, by Biber et al. (1999), meaning that processing would be affected by frequency of use in a continuous rather than a step-wise manner (Arnon and Snider 2010). If there is such a threshold, its value remains to be empirically determined (Nordquist 2009).

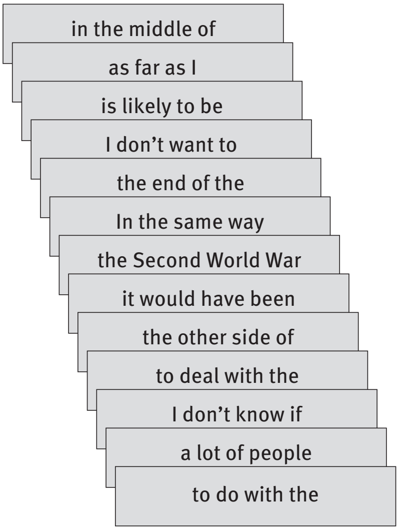

Figure 3: Four-word sequence production. Participants are asked to say out loud the four-word sequences that appear on the screen.

of N-grams of various lengths (including quadgrams). In the production onset time analysis, log probabilities of use were the most important variable family and accounted for 0.93% of the total variability in the data, closely followed by mutual information values (0.85%) and finally by frequencies of occurrence (0.21%). This suggests that the main process underlying recognition was one of competition (Marslen-Wilson 1995) between a bigram, trigram, or quadgram and other N-grams similar to them (i.e., *at the age of* competing with *at the age when*, *at the age where*, . . . for selection; see Section 6.1 for more details).

In the production duration analysis, however, frequencies of occurrence were by far the most important, accounting for 1.11% of the total variability in the data, and was followed by log probabilities of use (0.05%). Surprisingly, mutual information values did not have any explanatory power at this stage. What is important during production proper is the number of times one has accessed/produced a linguistic item. The neuromotor routines that instantiate a sequence’s phonetic form become more fluent with repetition resulting in reduction and coarticulation (Bybee 2001, 2006, and references cited therein).

In brief, these results suggest not only that multi-word sequences are stored in the mental lexicon, but also that they are processed in parallel as wholes and as parts. This brings further support to the existence of a redundant, inflationist lexicon.

4.2 Four-word sequence recall

Tremblay and Baayen (2010) ran a four-word sequence recall experiment using the same stimuli as in Tremblay and Tucker's (2011) production task. Participants were presented with a series of six four-word sequences and asked to remember them for subsequent recall (there were 72 blocks for a total of 432 four-word

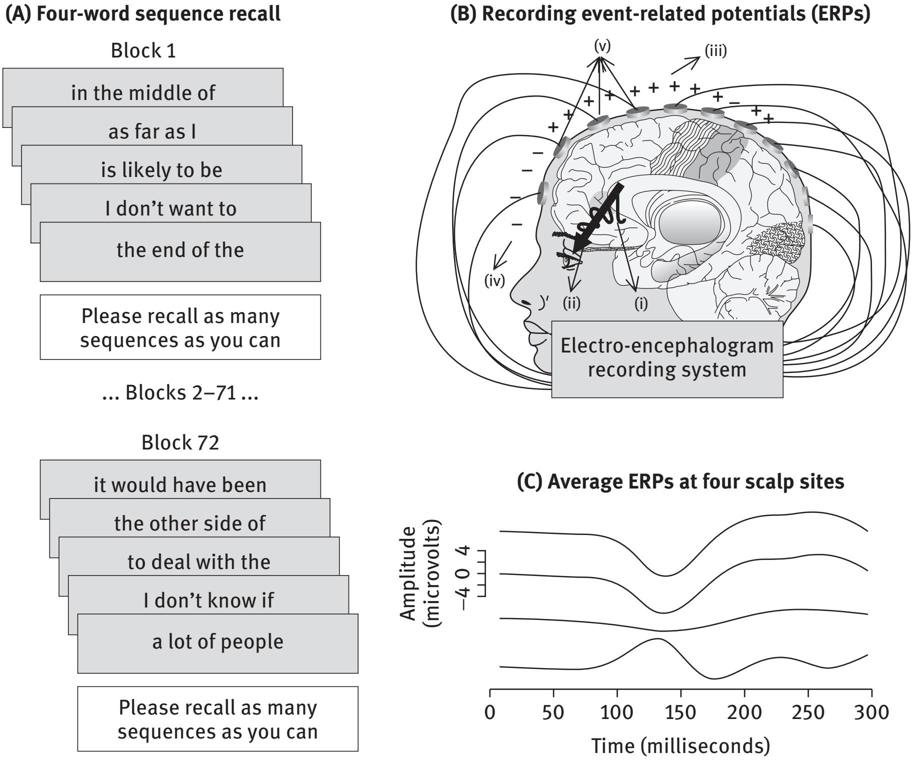

Figure 4: Four-word sequence recall with event-related brain potential recordings. (A) Participants were shown 432 four-word sequences six at a time (72 blocks) and asked to recall as many as they could. (B) While participants were performing the task, event-related brain potentials were recorded from their scalp. This portion of the experiment is discussed in Section Holistic retrieval or speeded composition?. (i) A group of neurons in cortex is firing synchronously. (ii) Each neuron produces a small electrical potential. These small potentials sum together to create a bigger potential (an ERP) that can be recorded from the scalp. (iii) The positive end of the potential. (iv) The negative end of the potential. (v) Electrodes placed on a participant's scalp to record ERPs. (C) From top to bottom: ERPs recorded at Fz (midline front), Cz (midline central), Pz (midline parietal), and Oz (midline occipital). The *x*-axis is time in milliseconds and the *y*-axis is microvolts.

sequences; see Figure 4A). While participants were performing the task, event-related brain potentials (ERPs, also known as electro-encephalograms or EEGs) were recorded from their scalp (see Figure 4B). We will discuss the ERP portion of the experiment in Section 5.

Statistical analyses revealed that the frequency of use of the third and fourth words of a sequence, of the first and second trigram of a sequence, and of the log probability of occurrence of the sequence as a whole affected the probability with which that sequence would be correctly recalled. Most interestingly, the quadgram probability effect varied according to whether the four-word sequence was a phrase (e.g., *in the same way*, *a lot of people*, *I don't want to*), or not (e.g., *in the middle of*, *to deal with the*, *I don't know if*).[3] Although the log probability effect was present whether a sequence was a stand-alone phrase or not (the higher a sequence's probability of occurrence, the better one's chances of correctly recalling it), the facilitatory effect was much stronger for stand-alone phrases.

In sum, given that whole-form frequency effects are commonly assumed to reflect holistic processing (e.g. Baayen, 2007, Bybee, 2007, Janssen, Bi, & Caramazza, 2008, Cholin, 2008), the results summarized here support the Cognitive concept of a redundant, inflationist lexicon where four-word sequences *as well as* their component parts are stored as wholes in the lexicon and processed in parallel.

5 Holistic retrieval or speeded composition?

As just mentioned above, whole-form frequency effects have been assumed to be the hallmark of holistic processing. Nevertheless, it is possible that such effects index speeded rule-based composition instead. Numerous studies have demonstrated syntactic priming (e.g. Grompel and Pickering 2007; Bernolet and Hartsuiker 2010; Haskell et al. 2010, Christianson et al. 2010), suggesting that Generative processes may be speeded by repetition. Following this line of thought, frequently-generated multi-word sequences may be (de)composed more quickly because

[3] Phrases are defined as having no missing material (they are complete). For instance, *in the same way* is a phrase because it can be inserted in a sentence as is, without adding anything to it: *[It can be done] [in the same way]*. However, *in the middle of* is not a phrase because linguistic material should be appended for it to be complete: *[He was walking] [in the middle of + the street]*. Sequences such as *I don't want to* are considered to be phrases because they are complete and can stand alone. For example, *JOHN: Do you want to go to the fair tonight? MARY: I don't want to.*

the application of compositional rules is more practiced. Therefore, the mental lexicon may still be an atomic one even if whole-form frequency is found to affect the processing of regular, non-idiomatic multi-word sequences. The present state of our knowledge, based largely on behavioural reaction-time studies, does not enable us to adequately discriminate between the two competing hypotheses. Given that whole-form frequency effects have been and remain at the centre of the "holistic processing versus on-line rule-based computation" debate, it is crucial to gain a better understanding of the exact nature of these effects. The only study we are aware of that investigated this specific issue is reported in Tremblay (2009) and Tremblay and Baayen (2010). Here the ERP component of the chunk recall experiment described above is discussed.

Event-related brain potentials are recordings of the electrical potentials generated by neurons. They are the means by which neurons communicate with one another. The potentials generated by single neurons are too small to be recorded at the scalp. However, if a large group of neurons start firing synchronously (see [i] in Figure 4B), the potentials generated by each individual neuron sum to a much bigger one (see [ii] in Figure 4B). Such a potential is called an event-related potential (ERP) because it is tied to the onset of a particular event, for instance the appearance of a four-word sequence on a computer screen. The ERP, which has a positive and a negative end (see [iii] and [iv] in Figure 4B), can be recorded from electrodes placed on the scalp (see [v] in Figure 4B) and subsequently analyzed. The ERPs look like the ones shown in Figure 4C, where the *x*-axis is time in milliseconds (a millisecond is a thousandth of a second) and the *y*-axis is microvolts (a microvolt is a millionth of a volt). The lines represent the average ERPs recorded at four midline positions on the scalp from the front of the head (Fz at the top of the figure) to its back (Oz at the bottom of the figure).

After analyzing the ERPs recorded while participants were performing the recall task, we found that the probability of use of a four-word sequence affected ERP amplitudes at frontal and posterior scalp sites 110–150 ms after a sequence appeared on the screen, which is thought to be the fastest *single words* can be accessed (e.g. Sereno et al. 1998; Hauk and Pulvermuller 2004; Penolazzi et al. 2007). At frontal sites (the front of the head), more negative ERP amplitudes were associated with greater quadgram probabilities. At posterior sites (the back of the head) ERP amplitudes were greatest for quadgrams with a lower probability of use and decreased rapidly until there was little or no activity (a flat line) from mid to higher probability values.

The fact that four-word sequences as wholes affected brain activity as early as single words have in previous studies suggests two things. First, at some stage of processing whole-form frequency and probability of occurrence may index holistic retrieval rather than speeded rule-based composition. Second, at least some regular, non-idiomatic four-word sequences are equivalent to single words in that they are stored as wholes in the mental lexicon.

6 Implications

The studies summarized here show that speakers are sensitive not only to the (continuous) frequency/probability of use of single words but also of compositional bigrams (*I really*), trigrams (*I really like*), and quadgrams (*I really like it*). They clearly support the Cognitive view of a redundant, inflationist lexicon that is comprised of a myriad of more or less entrenched linguistic units of different lengths. Such exemplars can be taken to be the basis on which emerge more and more abstract "frames" (Cienki 2007), "constructions" (Goldberg 1995; Croft 2007), "patterns" (Hunston and Francis 2000), "morpheme equivalent unit" (Wray 2008), or "lexical bundles" (Biber et al. 1999), such as for example *I __ like it*, *I __ __ it*, *__ really like it*, *__ __ like it*, *I really like __*, *I really __ __*, and *Subject + Verb + Object*.

6.1 Entrenchment of an N-gram

According to Langacker (1987: 59, 100) and Schmid (2007: 119), frequency is a measure of progressive (i.e., continuous) entrenchment of linguistic units of any degree of complexity. For Geeraerts et al. (1994), however, the entrenchment of a linguistic unit does not result from the number of times it was used in general, but rather from how many times it was used *relative* to other linguistic units related to it. Based on the study results presented above, both views are correct.

Before explaining why both views are correct, it is worth clarifying what the terms "frequency of use" and "probability of use" actually mean. While frequency of use measures the salience of a particular N-gram relative to everything else, probability of use is a measure of the salience of an N-gram relative to other N-grams of the same family. For example, the trigram *I think that's* can be thought of as a family of quadgrams which is comprised of *I think that's a*, *I think that's what*, *I think that's the*, *I think that's why*, and so forth.

In the studies summarized above that investigated N-gram lexical access (i.e. the self-paced reading tasks, the sentence and word recall experiments, the chunk recall experiment, as well as the pre-production stage of the speech production study), the probability of use of an N-gram emerged as the most important predictor. However, in the production stage of the speech production experiment frequency of use was the most important one. As Tremblay and Tucker (2011: 319–320) write, this may indicate that the main process underlying lexical access is one of competition between N-grams and their family members (Marslen-Wilson 1995), whereas sheer repetition underlies the entrenchment of the neuromotor routines that instantiate a sequence's phonetic form (Bybee 2001, 2006).

In a nutshell, the exact definition of the term "entrenchment" is dependent on the situation at hand, that is, whether one is accessing a linguistic unit (in this case, entrenchment should be defined as per Geeraerts et al. 1994) or whether one is producing it (where entrenchment should be construed as per Langacker 1987 and Schmid 2007).

6.2 Opacity of an entrenched N-gram

We usually think of an entrenched N-gram as being an un-analyzable, opaque unit. However, as Langacker (1987: 59) notes, "When a complex structure coalesces into a unit, its sub-parts do not thereby cease to exist or be identifiable as substructures . . . Its components do become less salient, however, precisely because the speaker no longer has to attend to them individually". The results reported here clearly support Langacker's idea. Indeed, the frequency/probability of use of quadgrams *as well as* their component parts (i.e., unigrams, bigrams, and trigrams) affected language processing in the chunk recall experiment and the speech production task.

The next step in better understanding the entrenchment and opacity (or lack thereof) of N-grams will be to take a closer look at the complex interactions that arise between N-grams of various lengths. One could hypothesize, along the lines of Langacker (1987: 59), that as the frequency/probability of use of an N-gram increases, the effects on language processing of its component parts will decrease. Nevertheless, these effects will not disappear entirely even when the frequency/probability of use of the "parent" N-gram is extremely high.

6.3 Threshold for N-gram holistic storage

An empirically based frequency/probability threshold for the holistic storage of an N-gram remains to be determined (Nordquist 2009). However, the results of the chunk recall experiment with electro-encephalogram recordings (EEG) provide a possible answer to this question, at least for stand-alone, phrasal quadgrams (i.e., *in the United States*, *they don't have to*, *she was going to*, and *he shook his head* versus non-phrasal ones such as *but there is no*, *the result of a*, and *I don't think it's*). The probability of occurrence of such quadgrams affected

ERP amplitudes recorded from electrodes placed at the back of the head 110–150 milliseconds after they were shown on a computer screen (Tremblay and Baayen 2010). In and of themselves, these results support the view that (at least some aspect of) quadgrams are stored and retrieved as wholes. This conclusion is further supported by the fact that (1) these ERPs are thought to originate from the ventral extra-striate cortex of the fusiform gyrus (Di Russo et al. 2002), and (2) functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) studies have reported activation in this cortical region during word, object, and face presentations, which diminished with repeated presentations (Rossion et al. 2001, and references cited therein). These latter observations, which mirror the results from the chunk recall experiment with EEG recordings, are generally attributed to "better (or faster) performance at processing these stimuli and could be taken to be the neural correlates of perceptual priming or implicit memory processing . . . In other words, these deactivations reflect a facilitation in neural computations when the same information is processed again" (Rossion et al. 2001: 1027).

More importantly here, the probability of occurrence effect on ERP amplitudes recorded at the back of the head were not a straight line, but rather U-shaped. For quadgrams with a (log) probability of occurrence between –7 (lowest) and –2 (mid), posterior ERP amplitudes decreased as the probability of occurrence increased. However, for quadgrams with a probability of occurrence between –2 (mid) and 4 (highest), posterior ERP amplitudes remained the same (i.e., did not increase or decrease). In short, it is conceivable that the threshold for holistic storage of phrasal quadgrams is a (log) probability of occurrence of –2.

7 Conclusion

In line with Cognitive models of language, the results of the studies described above point to the existence of a redundant, inflationist lexicon where non-idiomatic, regular sequences of words are stored *as well as* the shorter sequences contained within them. Given such a lexicon, what would the operations performed by the computational system look like? There should at least be two basic operations: Appending a pre-fabricated sequence to another one (e.g., *[He shook his head]* + *[for the first time]* → *He shook his head for the first time*), and filling in open slots with the appropriate lexical material (e.g., *[___ shook ___ head]* + *[she]* + *[her]* → *She shook her head*). There may be a third one namely, the deletion of some overlapping portions of two sequences that are to be appended to one another (e.g., *The problem is* + *is that the* → *The problem is that the*). Evidence that this third operation may exist comes from speech production errors such as *The problem* is is *that the* . . . In this particular case, it is as though the speaker retrieved the sequences *the problem is* and *is that the* as wholes and appended them to one another but "forgot", so to speak, to delete one of the two *is*-s. In the *Contemporary Corpus of American English* there are about 1000 such instances and some (very educated) native speakers of English I know make this error all the time. In fact, they say *the problem is is that the* so often that this sequence may have become one holistic unit stored in their lexicon.

A redundant lexicon entails that there is, at least in principle, more than one way to create one and the same utterance. The sentence *She shook her head for the first time* might be generated by retrieving a sequence with two open slots, filling them with the appropriate lexical material, and appending the resulting sequence to another four-word sequence (*[___shook___head]* + *[she]* + *[her]* → *[She shook her head]* + *[for the first time]* → *She shook her head for the first time*) or it may be composed in many other ways (e.g., *[She shook]* + *[her head]* + *[for the]* + *[first time]*, *[She]* + *[___ shook her]* + *[head]* + *[for the first time]*, and so forth). The question then is how can we know how an utterance was generated?

It is conceivable that the utterance formation process follows the path of least resistance. The most efficient path might be one where speakers retrieve the longest possible sequences that are the most easily accessible and assemble them together. Sometimes, it may be that a whole utterance is retrieved and used without further ado. Other times, retrieving two or more overlapping sequences and deleting the overlapping portions might be more efficient. Then again, one may wonder whether appending and deleting is more cost effective than filling out open slots. Furthermore, is retrieving longer sequences from the lexicon more resource intensive than retrieving smaller units? Is it more costly to perform operations on longer sequences than on shorter ones? Is it more effortful to use the storage system than the computational system? Moreover, can the path of least resistance be defined in the same way for every speaker? That is, does everyone rely to the same extent on the storage and computational systems or do some rely more heavily on the storage system while others on the computational system? Finally, can the formation of one specific utterance by one specific individual vary as a function of external factors such as time of day, how many margaritas one has had, and/or how well one has slept the night before?

To the best of my knowledge, these questions remain unanswered. Answering them, however, will bring us one step closer to figuring out exactly how we use language. This information will ultimately enable us to refine existing models of language and/or to develop new, cognitively plausible ones.

Dalhousie University, Canada

Acknowledgements

The studies reported in this paper were partly supported by a Major Collaborative Research Initiative Grant and a Doctoral Fellowship from the Social Sciences and Humanities Research Council of Canada (SSHRC), as well as a grant from the National Sciences and Engineering Research Council of Canada (NSERC). This paper was presented in part at an organized session of the *2011 Linguistic Society of America Annual Meeting* entitled *Empirical evaluations of usage-based constructionist models of language representation and processing*. I wish to thank past and present members of the Centre for Comparative Psycholinguistics, of Chris Westbury's lab, and of Benjamin V. Tucker's lab, all at the University of Alberta, without whom the studies reported here would not have been possible. I also wish to thank specifically R. Harald Baayen, Gary Libben, Taswar Bhatti, Patrick Bolger, Georgie Columbus, Perter Hendrix, Aaron Newman, and John Newman.

References

Alario, F.-Xavier., Albert Costa & Alfonso Caramazza. 2002. Frequency effects in noun phrase production: Implications for models of lexical access. *Language and Cognitive Processes* 17(3). 299–319.

Arnon, Inbal & Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62. 67–82.

Baayen, R. Harald. 2007. Storage and computation in the mental lexicon. In Gonia Jarema & Gary Libben (eds.), *The mental lexicon: Core perspectives*, 81–104. Amsterdam: Elsevier.

Bernolet, Sarah & Robert J. Hartsuiker. 2010. Does verb bias modulate syntactic priming? *Cognition* 114(3). 455–461.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow and Essex: Pearson Education Ltd.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, Joan. 2007. From usage to grammar: The minds response to repetition. *Language* 82(4). 711–733.

Bybee, Joan & Paul Hopper. 2001. *Frequency and the emergence of linguistic structure*. Amsterdam and Philadelphia: John Benjamins.

Bybee, Joan & Joanne Scheibman. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics* 37(4). 575–596.

Caramazza, Alfonso. 1997. How many levels of processing are there in lexical access. *Cognitive Neuropsychology* 14. 177–208.

Cholin, Joana. 2008. The mental syllabary in speech production: An integration of different approaches and domains. *Aphasiology* 22(11). 1127–1141.

Chomsky, Noam. 1993. A minimalist program for linguistic theory. In Ken Hale & Samuel J. Keyser (eds.), *The view from building 20: Essays in linguistics in honor of Sylvain Bromberger,* 1–52. Cambridge, MA: MIT Press.

Christianson, Kiel, Steven G. Luke & Fernanda Ferreira. 2010. Effects of plausibility on structural priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36(2). 538–544.

Cienki, Alan. 2007. Frames, idealized cognitive models, and domains. In Dirk Geeraerts & Hubert Cuykens (eds.), *The Oxford handbook of cognitive linguistics*, 170–187. Oxford: Oxford University Press.

Croft, William. 2007. Construction Grammar. In Dirk Geeraerts & Hubert Cuykens (eds.), *The Oxford handbook of cognitive linguistics*, 463–508. Oxford: Oxford University Press.

Davies, Mark. 2004. *BYU-BNC: The British National Corpus.* http://corpus.byu.edu/bnc.

Davies, Mark. 2008. *The Corpus of Contemporary American English (COCA): 400+ Million Words, 1990-Present.* http://www.americancorpus.org.

Dell, Gary S. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93(3). 283–321.

Di Russo, Francesco, Antigona Martinez, Martin I. Sereno, Sabrina Pitzalis, & Steven A. Hillyard. 2002. Cortical sources of the early components of the visual evoked potential. *Human Brain Mapping* 15. 95–111.

Ellis, Nick C. & Rita Simpson-Vlach. 2009. Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory* 5(1). 61–78.

Embick, David, Alec Marantz, Yasushi Miyashita, Wayne O'Neil & Kuniyoshi L. Sakai. 2000. A syntactic specialization for Broca's area. *Proceedings of the National Academy of Sciences of the United States of America* 97(11). 6150–6154.

Geeraerts, Dirk, Stefan Grondelaers & Peter Bakema. 1994. *The structure of lexical variation: Meaning, naming, and context.* Berlin: Mouton de Gruyter.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, Adele E. 2009. The nature of generalization in language. *Cognitive Linguistics* 20. 93–127.

Grodzinsky, Yosef & Angela D. Friederici. 2006. Neuroimaging of syntax and syntactic processing. *Current Opinion in Neurobiology* 16(2). 240–246.

Grompel, Roger P.G. van & Martin J. Pickering. 2007. Syntactic parsing. In M. Gareth Gaskell (ed.), *The Oxford handbook of psycholinguistics*, 289–307. Oxford: Oxford University Press.

Halle, Morris & Alec Marantz. 1993. Distributed morphology and the pieces of inflection. In Ken Hale & Samuel J. Keyser (eds.), *The view from building 20: Essays in Linguistics in Honor of Sylvain Bromberger,* 111–176. Cambridge, MA: MIT Press.

Hand, Christopher. J., Sebastian Miellet, Patrick J. O'Donnell & Sara C. Sereno. 2010. The frequency-predictability interaction in reading: It depends where you're coming from. *Journal of Experimental Psychology-human Perception and Performance* 36(5). 1294–1313.

Haskell, Todd R., Robert Thornton & Maryellen C. MacDonald. 2010. Experience and grammatical agreement: Statistical learning shapes number agreement production. *Cognition* 114(2). 151–164.

Hauk, Olaf & Friedemann Pulvermuller. 2004. Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology* 115. 1090–1103.

Hunston, Susan & Gill Francis. 2000. *Pattern grammar: A Corpus-driven approach to the lexical grammar of English*. Amsterdam and Philadelphia: John Benjamins.

Jackendoff, Ray S. 2002. *Foundations of language*. Oxford: Oxford University Press.

Janssen, Niels, Yanchao Bi & Alfonso Caramazza. 2008. A tale of two frequencies: Determining the speed of lexical access for Mandarin Chinese and English compounds. *Language and Cognitive Processes* 23(7/8). 1191–1223.

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

Jiang, Nan & Tatiana M. Nekrasova. 2007. The processing of formulaic sequences by second language speakers. *The Modern Language Journal* 91(3). 433–445.

Joanisse, Marc F. & Mark S. Seidenberg. 2005. Imaging the past: Neural activation in frontal and temporal regions during regular and irregular past-tense processing. *Cognitive, Affective & Behavioral Neuroscience* 5(3). 282–96.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar. Vol. 1, theoretical prerequisites.* Stanford: Stanford University Press.

Langacker, Ronald W. 1991. *Foundations of cognitive grammar. Vol. 2, descriptive application.* Stanford: Stanford University Press.

Levelt, Willem J. 1989. *Speaking: From intention to articulation*. Cambridge, MA: The MIT Press.

Levelt, Willem J. 1992. Accessing words in speech production: Stages, processes and representations. *Cognition* 42. 1–22.

Levelt, Willem J., Ardi Roelofs & Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22. 1–38.

Marantz, Alec. 1995. The minimalist program. In Gert Webelhuth (ed.), *Government and binding theory and the minimalist program,* 349–382. Cambridge, Mass.: Blackwell.

Marslen-Wilson, William. 1995. Activation, competition, and frequency in lexical access. In Gerry T. Altman (ed.), *Cognitive models of speech processing,* 148–172. Cambridge, MA: MIT Press.

Marslen-Wilson, William. 2007. Morphological processes in language comprehension. In M. Gareth Gaskell (ed.), *The Oxford handbook of psycholinguistics*, 175–193. Oxford: Oxford University Press.

McClelland, James L. 1987. The case for interactionism in language processing. *Attention and Performance* 12. 3–36.

Miller, George A. 1956. The magical number seven, plus or minus two. *Psychological Review* 63. 81–97.

Newman, Aaron, Michael T. Ullman, Roumyana Pancheva, Diane L. Waligura & Helen J. Neville. 2007. An ERP study of regular and irregular English past tense inflection. *Neuroimage* 34. 435–445.

Nordquist, Dawn. 2009. Investigating elicited data from a usage-based perspective. *Corpus Linguistics and Linguistic Theory* 5. 105–130.

Penolazzi, Barbara, Olaf Hauk & Friedemann Pulvermuller. 2007. Early semantic context integration and lexical access as revealed by event-related brain potentials. *Biological Psychology* 72. 373–388.

Pinker, Steven & Michael T. Ullman. 2002. Combination and structure, not gradedness, is the issue: Reply to McClelland and Patterson. *Trends in the Cognitive Sciences* 6(11). 472–474.

Rossion, Bruno, Christine Schiltz, Laurence Robaye, David Pirenne & Marc Crommelinck. 2001. How does the brain discriminate familiar and unfamiliar faces? A pet study of face categorical perception. *Journal of Cognitive Neuroscience* 13. 1019–1034.

Savin, Harris B. & Ellen Perchonock. 1965. Grammatical structures and the immediate recall of English sentences. *Journal of Verbal Learning and Verbal Behaviour* 4. 384–353.

Schmid, H. Jorg. 2007. Entrenchement, salience, and basic levels. In Dirk Geeraerts & Hubert Cuykens (eds.), *The Oxford handbook of cognitive linguistics*, 117–138. Oxford: Oxford University Press.

Schmitt, Norbert. 2005. Formulaic language: Fixed and varied. *Estudios de Linguística Inglesa Aplicada* 6. 13–39.

Schmitt, Norbert, Sarah Grandage & Svenja Adolphs. 2004. Are corpus-derived recurrent clusters psycholinguistically valid? In Norbert Schmitt (ed.), *Formulaic sequences: Acquisition, processing and use*, 173–189. Amsterdam and Philadelphia: John Benjamins.

Sereno, Sara C., Keith Rayner & Michael I. Posner. 1998. Establishing a time-line of word recognition: Evidence from eye movements and event-related potentials. *NeuroReport* 9(10). 2195–2200.

Tremblay, Antoine. 2009. *Processing advantages of lexical bundles: Evidence from self-paced reading, word and sentence recall, and free recall with event-related brain potential recordings.* Edmonton, Canada: University of Alberta PhD dissertation.

Tremblay, Antoine & R. Harald Baayen. 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 David Wood (ed.), *Perspectives on formulaic language: Acquisition and communication*, 151–173. London and New York: Continuum.

Tremblay, Antoine, Bruce L. Derwing, Gary Libben & Chris Westbury. 2011. Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning* 61(2). 569–613.

Tremblay, Antoine & Benjamin V. Tucker. 2011. The effects of n-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon* 6(2). 302–324.

Underwood, Geoffrey, Nobert Schmitt & Adam Galpin. 2004. The eyes have it: An eye-movement study into the processing of formulaic sequences. In Norbert Schmitt (ed.), *Formulaic sequences: Acquisition, processing and use*, 153–172. Amsterdam and Philadelphia: John Benjamins.

Weinert, Regina. 2010. Formulaicity and usage-based language: Linguistic, psycholinguistic and acquisitional manifestations. In David Wood (ed.), *Perspectives on formulaic language: Acquisition and communication*, 1–22. London and New York: Continuum.

Wray, Alison. 2002. *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.

Wray, Alison. 2008. *Formulaic language: Pushing the boundaries.* Oxford: Oxford University Press.