

Empirical Evidence for an Inflationist Lexicon

Antoine Tremblay (trea26@gmail.com)

Research Services, IWK Health Centre,
Goldbloom Research and Clinical Care Pavilion,
5850/5980 University Avenue, P.O. Box 9700,
Halifax, Nova Scotia, B3K 6R8, Canada

R. Harald Baayen (baayen@ualberta.ca)

Department of Linguistics, 4-55 Assiniboia Hall,
University of Alberta, Edmonton T6G 2E5, Canada

Bruce Derwing (bderwing@me.com)

Department of Linguistics, University of Alberta,
Edmonton, AB Canada T6G 2E7, Canada

Gary Libben (glibben@ucalgary.ca)

Associate Vice-President Research, University of Calgary,
Administration 100, 2500 University Drive N.W.,
Calgary T2N 1N4, Canada

Benjamin V. Tucker (bvtucker@ualberta.ca)

Department of Linguistics, University of Alberta,
4-32 Assiniboia Hall, Edmonton, Alberta T6G 2E7, Canada

Chris Westbury (chrisw@ualberta.ca)

Department of Psychology, University of Alberta,
P577 Biological Science Building, Edmonton,
Alberta, T6G 2E9, Canada

Abstract

Although generative and construction grammars both assume that some linguistic forms are stored/retrieved 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 construction grammarians frequency of use is an important factor determining whether a form is stored as a whole or (de)composed on-line. We present empirical evidence from self-paced reading, sentence recall, and chunk production experiments showing that speakers are sensitive to the frequency of use of regular, non-idiomatic multi-word sequences (*at the end of*), thus suggesting that they are stored/retrieved as wholes (favoring the constructionist view). However, these frequency effects could rather reflect speeded/practiced rule-based (de)composition. Results from our chunk recall experiment with event-related potential recordings suggest that (some aspects) of such multi-word sequences may be holistically stored/retrieved.

Keywords: Multi-word sequences; four- and five-word lexical bundles; self-paced reading; sentence recall; four-word sequence recall; electro-encephalogram; EEG; ERP; generalized additive modeling; GAM

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 constructionist grammars are two examples of dual-route mod-

els, 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 & 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*). For construction grammarians, however, (e.g., Bybee & Hopper, 2001; Goldberg, 2009; Sinclair, 1991; Wray, 2002, 2008) 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* 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 constructionist 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 constructionist lexicon incorporates redundancy and is thus inflationist.

Although there is evidence for both the generative (e.g., Embick, Marantz, Miyashita, O’Neil, & Sakai, 2000; Pinker & Ullman, 2002a; Grodzinsky & Friederici, 2006; Newman, Ullman, Pancheva, Waligura, & Neville, 2007; Marslen-Wilson, 2007) and the constructionist views (e.g., Arnon & Snider, 2010; Bybee & Scheibman, 1999; Bybee & Hopper, 2001; Wray, 2002; Joanisse & Seidenberg, 2005; Bybee, 2007; Wray, 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.

Lexical Bundles and Self-paced Reading

In Tremblay, Derwing, Libben, and Westbury (in press), we investigated the idea put forth by Biber et al. (1999) that lexical bundles such as *in the middle of the* are stored and processed holistically. In three self-paced reading experiments we compared sentences containing lexical bundles (e.g., *in the middle of the*) and matched control multi-word sequences (e.g., *in the front of the*). Four-word 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). Note that lexical bundles may span traditional syntactic boundaries. Frequency counts were taken from the *BYU: British National Corpus* (BNC; Davies, 2004). Lexical bundles and control sequences differed in one word, which, in target lexical bundle sequences, was never shorter or more frequent than in control sequences. Lexical bundles and control 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 the first experiment, sentences were presented one word at a time. Participants navigated through the sentences by pressing the space bar. In the second experiment, sentences were presented portion-by-portion. The first portion, which did not contain the target or control sequence, appeared in the center of a computer screen; the remaining two portions appeared after participants pressed the space bar (once for each portion). In the third experiment, sentences were presented as wholes; participants had to press the space bar to see the next sentence. Reading times were measured from the moment a word/portion/sentence appeared on the screen to the time a participant made a button press. In all three experiments, lexical bundles and sentences containing lexical bundles were read faster than the control multi-word sequences thus lending support to hypothesis put forth by Biber et al. (1999).

Lexical Bundles and Sentence Recall

The lexical bundle facilitatory effect was replicated in two follow-up word and sentence recall experiments (in the auditory and visual modalities). In these experiments, sentences containing lexical bundles and control non-lexical bundle sequences were presented to participants. Each sentence was followed by a series of six monomorphemic words (of equal length and frequency of use, e.g., *date, dice, male, page, pool, tape*). Participants were asked to recall the sentences

and as many single words as they could remember. In both the auditory and visual modalities, sentences containing lexical bundles were more accurately recalled than those that did not. It was 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 stored and retrieved as wholes as put forth by Biber et al. (1999).

Non-idiomatic Multi-word Sequences, Production, and Recall

In the latter five experiments, we strictly looked at the effects of the frequency of occurrence of four- and five-word sequences 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, which would lend further support to the inflationist lexicon concept. Moreover, there may be interactions between single word, bigram, trigram, and quadgram frequency, which would indicate that these variables share the same processing stage (Hand, Mielliet, O’Donnell, & Sereno, 2010). Furthermore, probabilistic measures such as mutual information (the association strength between two or more words) and/or the probability of occurrence (the probability of a word occurring given a previous context of a certain length) may also affect multi-word sequence processing and may even trump frequency of use (as in, e.g., Ellis & Simpson-Vlach, 2009). Finally, there may not be a threshold for holistic storage meaning that frequency effects affect processing in a continuous rather than a step-wise manner (Arnon & Snider, 2010).¹ Tremblay and Tucker (in preparation) investigated these issues in the context of a four-word sequence production task. Four hundred and thirty two four-word sequences were presented to participants one at a time (e.g., *at the end of, I don’t think that, in the United States*). Participants were simply asked to say them aloud. Production onset latencies were measured from the onset of presentation to the moment participants began production; production durations were also measured. Whole-sequence (quadgram) frequency counts, taken from the *Contemporary Corpus of American English* (COCA; Davies, 2008), ranged roughly from 0.03 to 85 occurrences per million words. Single word, bigram, and trigram frequencies were also obtained; bigram, trigram, and quadgram mutual information and log probability of occurrence values were computed. Our analyses revealed numerous main effects as well as interactions between single-word frequencies and the frequency of occurrence, probability of occurrence, and mutual information of N-grams of various lengths (including quadgrams). In Tremblay and Baayen (2010)’s four-word sequence recall experiment, where the same stimuli as in the Tremblay and Tucker (in preparation) paper were used, a number of significant main effects and interactions between probabilistic

¹If there is such a threshold, its value remains to be empirically determined (Nordquist, 2009).

measures of N-grams of various lengths were also found (including quadgrams; see below for more details about the experiment).

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 presented here support the constructionist concept of a redundant, inflationist lexicon where four-word sequences *as well as* smaller sequences and single words would all be stored as wholes in the lexicon (e.g., *at-the-end-of, at-the-end, the-end-of, at-the, the-end, end-of, at, the, end, of*).

Holistic Storage / Retrieval or Speeded / Practiced (De)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/practiced rule-based (de)composition instead. Numerous studies have demonstrated syntactic priming (e.g., Gropel & Pickering, 2007; Bernolet & Hartsuiker, 2010; Haskell, Thornton, & MacDonald, 2010; Christianson, Luke, & Ferreira, 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 the application of compositional rules is more practiced, either for particular syntactic structures or possibly for the combination of specific multi-word sequences. 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 for instance. 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 center 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 Tremblay and Baayen (2010).

In this study, participants were presented with 72 series of six four-word sequences (sequences were presented as wholes). After each series, they were asked to recall as many sequences as they could. While they were performing the task, electro-encephalogram recordings were made (EEGs, which are recordings of the brain’s electrical activity). Using generalized additive modeling (GAM; see Tremblay, 2009; Hendrix, 2009; Tremblay & Baayen, 2010, for more details) we found that the probability of occurrence of a four-word sequence affected EEG amplitudes at frontal sites 110–150 ms after a sequence appeared on the screen (after controlling for single-word and N-gram probabilistic variables as well as other confounding factors), which is thought to be the fastest *single words* can be accessed (e.g., Sereno, Rayner, & Posner, 1998; Hauk & Pulvermuller, 2004; Penolazzi, Hauk, & Pulvermuller, 2007). Although this finding suggests that (at

least at some stage of processing) whole-form frequency may index holistic storage and retrieval rather than speeded rule-based composition, the jury is still out on this issue.

Acknowledgments

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). We wish to thank past and present members of the Center 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. We also wish to thank specifically Taswar Bhatti, Patrick Bolger, Georgie Columbus, Perter Hendrix, Aaron Newman, and John Newman.

References

- Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*, 67–82.
- Baayen, R. H. (2007). Storage and computation in the mental lexicon. In G. Jarema & G. Libben (Eds.), *The mental lexicon: Core perspectives* (p. (in press)). Amsterdam: Elsevier.
- Bernolet, S., & Hartsuiker, R. J. (2010). Does verb bias modulate syntactic priming? *Cognition, 114*, 455–461.
- Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English*. Harlow and Essex: Pearson Education Ltd.
- Bybee, J. (2007). From usage to grammar: The mind’s response to repetition. *Language, 82*(4), 711–733.
- Bybee, J., & Hopper, P. (2001). *Frequency and the emergence of linguistic structure*. Amsterdam and Philadelphia: John Benjamins.
- Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: the reduction of *don’t* in English. *Linguistics, 34*, 575–596.
- Cholin, J. (2008). The mental syllabary in speech production: An integration of different approaches and domains. *Aphasiology, 22*(11), 1127–1141.
- Chomsky, N. (1993). A minimalist program for linguistic theory. In K. L. Hale & S. J. Keyser (Eds.), *The view from building 20: Essays in linguistics in honor of sylvain bromberger*. (pp. 1–52). Cambridge, MA: MIT Press.
- Christianson, K., Luke, S. G., & Ferreira, F. (2010). Effects of Plausibility on Structural Priming. *Journal of Experimental Psychology-learning Memory and Cognition, 36*(2), 538–544.
- Davies, M. (2004). *Byu-bnc: The british national corpus*. Available online at <http://corpus.byu.edu/bnc>.
- Davies, M. (2008). *The corpus of contemporary american english (coca): 400+ million words, 1990-present*. Available online at <http://www.americancorpus.org>.

- Ellis, N. C., & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory*, 5, 61–78.
- Embick, D., Marantz, a., Miyashita, Y., O’Neil, W., & Sakai, K. L. (2000). A syntactic specialization for Broca’s area. *Proceedings of the National Academy of Sciences of the United States of America*, 97(11), 6150–4.
- Goldberg, A. (2009). The nature of generalization in language. *Cognitive Linguistics*, 20, 93–127.
- Grodzinsky, Y., & Friederici, A. D. (2006). Neuroimaging of syntax and syntactic processing. *Current opinion in neurobiology*, 16(2), 240–6.
- Grompel, R. P. van, & Pickering, M. J. (2007). Syntactic parsing. In M. G. Gaskell (Ed.), *The oxford handbook of psycholinguistics*. (pp. 289–307). Oxford: Oxford University Press.
- Halle, M., & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In K. Hale & S. J. Keyser (Eds.), *The view from building 20: Essays in linguistics in honor of Sylvain Bromberger* (Vol. 24, p. 111-176). Cambridge, Mass: MIT Press.
- Hand, C. J., Mielliet, S., O’Donnell, P. J., & Sereno, S. C. (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, T. R., Thornton, R., & MacDonald, M. C. (2010, FEB). Experience and grammatical agreement: Statistical learning shapes number agreement production. *Cognition*, 114(2), 151–164.
- Hauk, O., & Pulvermuller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, 115, 1090–1103.
- Hendrix, P. (2009). *Electrophysiological effects in language production: a picture naming study using generalized additive modeling*. MA dissertation, Radboud University, Nijmegen, the Netherlands.
- Jackendoff, R. (2002). *Foundations of language*. Oxford: Oxford University Press.
- Janssen, N., Bi, Y., & Caramazza, A. (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.
- Joanisse, M. F., & Seidenberg, M. S. (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.
- Marantz, A. (1995). The minimalist program. In G. Webelhuth (Ed.), *Government and binding theory and the minimalist program*. (pp. 349–382). Cambridge, Mass.: Blackwell.
- Marslen-Wilson, W. (2007). Morphological processes in language comprehension. In M. G. Gaskell (Ed.), *The oxford handbook of psycholinguistics* (pp. 175–193). Oxford: Oxford University Press.
- Newman, A., Ullman, M., Pancheva, R., Waligura, D. L., & Neville, H. J. (2007). An erp study of regular and irregular english past tense inflection. *NeuroImage*, 34, 435–445.
- Nordquist, D. (2009). Investigating elicited data from a usage-based perspective. *Corpus Linguistics and Linguistic Theory*, 5, 105–130.
- Penolazzi, B., Hauk, O., & Pulvermuller, F. (2007). Early semantic context integration and lexical access as revealed by event-related brain potentials. *Biological Psychology*, 72, 373–388.
- Pinker, S., & Ullman, M. (2002a). Combination and structure, not gradedness, is the issue: Reply to McClelland and Patterson. *Trends in the Cognitive Sciences*, 6(11), 472–474.
- Pinker, S., & Ullman, M. (2002b). The past and future of the past tense. *Trends in the Cognitive Sciences*, 6(11), 456–462.
- Pyllkanen, L., & Marantz, A. (2003). Tracking the time course of word recognition with MEG. *Trends in Cognitive Sciences*, 7, 187–189.
- Sereno, S., Rayner, K., & Posner, M. (1998). Establishing a time-line of word recognition: evidence from eye movements and event-related potentials. *NeuroReport*, 9(10), 2195–2200.
- Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Tremblay, A. (2009). *Processing advantages of lexical bundles: Evidence from self-paced reading, word ad sentence recall, and free recall with event-related brain potential recordings*. PhD dissertation, University of Alberta, Edmonton, Canada.
- Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and erp study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (p. 151-173). London and New York: Continuum.
- Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (in press). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, 61(3).
- Tremblay, A., & Tucker, B. V. (in preparation). What can the production of four-word sequences tell us about the mental lexicon? *The Mental Lexicon*.
- Weinert, R. (2010). Formulaicity and usage-based language: Linguistic, psycholinguistic and acquisitional manifestations. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 1–22). London and New York: Continuum.
- Wray, A. (2002). *Language and the lexicon*. Cambridge: Cambridge University Press.
- Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.