

More than words: Frequency effects for multi-word phrases

Inbal Arnon

Department of Linguistics, Stanford University,

Margaret Jacks Hall, Bldg 460,

Stanford CA 94305-2150

inbalar@stanford.edu

Neal Snider

Department of Brain and Cognitive Sciences, University of Rochester,

Meliora Hall, Box 270268,

University of Rochester,

Rochester, NY 14627-0268

nsnider@bcs.rochester.edu

Abstract

There is mounting evidence that language users are sensitive to distributional information at many grain-sizes. Much of this research has focused on the distributional properties of words, the units they consist of (morphemes, phonemes), and the syntactic structures they appear in (verb-categorization frames, syntactic constructions). In a series of studies we show that comprehenders are also sensitive to the frequencies of compositional four-word phrases (e.g., don't have to worry): more frequent phrases are processed faster. The effect is not reducible to the frequency of the individual words or substrings and is observed across the entire frequency range (for low, mid and high frequency phrases). Comprehenders seem to learn and store frequency information about multi-word phrases. These findings call for processing models that can capture and predict phrase frequency effects, and support accounts where linguistic knowledge consists of patterns of varying sizes and levels of abstraction.

Introduction

There is mounting evidence that language users are sensitive to distributional information at many grain-sizes: from that of sound combinations, through morphemes and single words, to syntactic constructions. Word recognition is affected (among other things) by the frequency of the word itself (Morton, 1969; see Monsell, 1991 for a review). Sentence comprehension is affected by a multitude of distributional factors, including the frequency of words (Rayner & Duffy, 1986); the frequency of words in specific syntactic structures (verb-subcategorization biases, Clifton, Frazier & Connine, 1984; MacDonald, Pearlmutter, & Seidenberg, 1994; Garnsey, Pearlmutter, Meyers & Lotocky, 1997); co-occurrence relations between verbs and specific arguments (Trueswell, Tanenhaus & Garnsey, 1994); as well as the overall frequency of syntactic structure (e.g. main clause vs. reduced relative, Frazier & Fodor, 1978). Production is also affected by the distributional properties of units of varying sizes. It is affected by word frequency (Jescheniak and Levelt, 1994), by the likelihood of a word given the previous one (Jurafsky et al, 2001), as well as the likelihood of the syntactic structure the word is part of (Gahl & Garnsey, 2004; Jaeger, 2006; Tily et al, 2009). Taken together, these findings show that language users are sensitive to detailed distributional information on many levels of linguistic analysis (see Ellis, 2002; Diessel, 2007 for reviews).

These findings have had implications for two distinct lines of research: one concerned with the processing of linguistic material, another with its mental representation. On the one hand, frequency effects have shaped and influenced models of processing. They led to the formulation of frequency-sensitive comprehension and production models that can account for the way different sources of information are

integrated in real-time processing. Such models include (but are not limited to) constraint-satisfaction and expectation-based models of comprehension (Hale, 2001; Jurafsky, 1996; 2003; MacDonald, 1994; MacDonald, Pearlmutter, & Seidenberg, 1994; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; McRae, Spivey-Knowlton & Tanenhaus, 1998). Similar frequency-sensitive models have been developed for production (e.g., Chang, Dell & Bock, 2006; Dell, 1986; Dell, Chang, and Griffin, 1999; Levy and Jaeger, 2007; Jaeger 2006). Uncovering the full range of distributional information that speakers are sensitive to becomes important for (a) developing adequate processing models, and (b) tackling the grain-size issue (Mitchell et al. 1995) that results from being able to estimate frequencies at multiple levels of linguistic analysis (e.g. syntactic construction, syntactic construction given a verb, syntactic construction given a verb and object, etc.). What is the relevant grain-size for a given calculation and how should different frequency measures be integrated? To be able to address this question, we need to know the full range of grain-sizes that language users attend to.

At the same time, frequency effects have influenced theories and models concerned with the way linguistic knowledge is learned and represented. There are a growing number of language models where linguistic units and categories are formed on the basis of experience. We group such models together under the label ‘emergentist models’, and include in this group usage-based approaches to grammar (Bybee, 1998; Goldberg, 1995, 2006; Langacker, 1987; Tomasello, 2003), connectionist models of learning and processing (e.g. Christiansen & Chater, 1999; Elman, 1990, 1991; MacWhinney, 1998; Rumelhart & McClelland, 1986; Seidenberg, 1994), and exemplar models of linguistic knowledge (Bod, 1998, 2006; Gahl & Yu, 2007; Goldinger, 1996;

Johnson, 1997; Pierrehumbert, 2001). In such models, experience plays an important role in the creation, entrenchment and processing of linguistic patterns. Grammatical structures emerge from, and are shaped by language use. Frequency of occurrence is seen as an index of linguistic experience. Specific models differ in many important respects (e.g. the use of symbolic or non-symbolic representations, see Bybee & McClelland, 2005). But they are similar in suggesting that all linguistic material is represented and processed in a similar way, and will be similarly affected by experience. In this context, frequency effects are interesting because they tell us about the linguistic units that speakers learn and represent.

Emergentist models stand in contrast to a words-and-rules approach to language (Pinker, 1991; Prince & Pinker, 1988; Pinker, 1999; Pinker & Ullman, 2002) where there is a clear distinction between the mental lexicon – an inventory of memorized forms, and the mental grammar – the rules or constraints used to combine them. This approach posits a distinction between the linguistic forms that are stored in the lexicon, and the ones that are computed by grammar. The two components of language are thought to be learned differently, to involve different cognitive abilities, and in some models, to be governed by different neural substrates (Ullman, 2001; Ullman et al. 2005). Frequency effects are interesting only as a way to distinguish between ‘stored’ and ‘computed’ forms: frequency is expected to affect the processing of stored forms but not computed ones (Pinker & Ullman, 2002).

Much of the research on frequency effects has focused on the distributional properties of words, the units they consist of (morphemes, phonemes), and the syntactic structures they appear in (verb-categorization frames, syntactic constructions). Less work

has focused on larger chunks of language (we review this literature in more detail in the next section). Here we ask whether processing is affected by the frequency of compositional multi-word phrases like ‘don’t have to worry’. By focusing on such units we can address, and tie together, two distinct questions.

From the perspective of processing models, looking at multi-word phrases fills an empirical gap. We know relatively little about the processing of such linguistic patterns. Yet such effects would expand our understanding of the units people attend to – multi-word phrases in addition to phonemes, morphemes, words and constructions,- and inform and limit the kind of models used to accommodate frequency effects. Word and bigram can be easily accommodated via links between words (or a non-symbolic representation of them) but frequency effects beyond the bigram (e.g. phrase-frequency effects) call for the representation of larger chains of relations (sequential information), not only between single words but also between word strings of varying sizes.

Looking for phrase-frequency effects is also interesting from a representational perspective. Under emergentist models, there isn’t a clear distinction between compositional multi-word phrases and smaller, more atomic linguistic patterns like words. Phrases can be represented by the same mechanism that represents words. For example, as independent units in an exemplar model (Bod, 1998; 2006) or as an attractor in connectionist models (Rodriguez, Willes & Elman, 1999). The processing of phrases, like that of words, should be affected by frequency. No such prediction is made under a words-and-rules model where compositional phrases are predicted to be computed and not stored. Since frequency effects are thought to be a property of memorized forms (Pinker & Ullman, 2002), compositional phrases are unlikely to exhibit frequency effects.

One way to contrast the two approaches is to look for phrase-frequency effects. We draw on a method used in the morphological literature to address a similar controversy regarding the representational status of regularly inflected words. Like compositional phrases, regularly inflected words (like *walked*) are expected to be generated in a words-and-rules approach (Pinker, 1999; Ullman & Pinker, 2002). To test this assumption, researchers ask whether the frequency of the inflected form (*walked*) is predictive of processing latencies independently of the frequency of its stem and all its inflectional variants (*walk, walks, walking, walked*). Such an effect is expected only if regularly inflected forms are represented as whole words. Many studies show whole-word frequency effects for regularly inflected words (Alegre & Gordon, 1999; Baayen et al. 1997, 2002; Stemberger & MacWhinney, 1988; Taft, 1979). These findings are taken as evidence for whole-form representation.

In a recent study, Bannard & Matthews (2008) extended this method to multi-word phrases. They aim to show that children store multi-word phrases. This is a crucial assumption for certain models of language development where grammatical knowledge is learned by abstracting over stored utterances (Tomasello, 2003; Abbot-Smith & Tomasello, 2006). To do so, they used a whole-form frequency manipulation with phrases. They compared the production of phrases that differed in phrase frequency but were matched on all other frequency measures. For example, *a drink of milk* is more frequent as a phrase than *a drink of tea* in British child-directed speech. But the two phrases are matched on substring frequency (*tea* is as frequent as *milk*, *of milk* is as frequent as *of tea*, and so on). They are also equally plausible. Any difference in performance has to reflect the properties of the phrase itself. Two and three-year-olds

were faster and better at repeating higher frequency phrases compared to lower frequency ones. Just as in the morphological literature, the authors took these effects to indicate whole-phrase representation.

The current study

In the current study we use a similar manipulation to investigate phrase-frequency effects in adult comprehension. There are now several motivations for looking at such phrases. The empirical findings of frequency at varying grain-sizes combined with the predictions of emergentist models predict phrase-frequency effects. But such effects have not been previously documented. At the same time, such effects allow us to contrast different views on the way linguistic knowledge is represented. Specifically, on whether there is a clear-cut qualitative distinction between compositional and simple forms (e.g. non-inflected words). Finding whole-form frequency effects for compositional phrases would argue against such a distinction. However, we do not want to argue that finding phrase-frequency effects implies that the phrases are stored as unanalyzed wholes. While this has been suggested for very frequent phrases (Bybee, 2002), it is not a claim we set out to investigate. Nor is it very likely given recent evidence that even idiomatic phrases, which are often thought to stored as unanalyzed wholes (Pinker, 1999), show evidence of internal structure (Konopcka & Bock, 2008).

We can roughly distinguish between three more detailed views on the representational status of multi-word phrases. A words-and-rules approach like that presented by Pinker and colleagues (Pinker, 1999; Pinker & Ullman, 2002), does not expect compositional phrases to be represented. Only non-compositional expressions,

like certain idioms, are expected to be stored. Such an approach does not predict phrase-frequency effects for compositional phrases.

A more nuanced position is presented in a frequency-threshold approach where phrases that are of sufficient frequency can attain independent representation as a way of making processing more efficient (Biber et al., 1999; Goldberg, 2006; Wray, 2002). Researchers differ on whether frequency is the only criterion for storage (as in the lexical bundles literature, Biber et al. 1999), or whether other factors also play a role in determining if a phrase is stored (e.g. compositionality in the case of Goldberg's Construction Grammar, 2006, or internal structure and context of use in Wray's study of formulaic language, 2002). A frequency-threshold approach maintains a distinction between phrases that are stored and ones that are not while allowing for more expressions to be stored in the lexicon.

No such distinction exists in what we will label a 'continuous' approach. In this emergentist framework, every instance of usage affects representation and processing. Compositional phrases are represented in the same way that simple words and non-compositional phrases are. The frequency of a phrase will influence its entrenchment and future processing (Bybee, 1998; Bybee, 2006; Bybee & Hopper, 2001; Bod et al. 2003). The difference between higher and lower frequency phrases is one of degree (the level of activation), and not of kind (stored vs. computed). This approach predicts frequency effects also for lower frequency phrases. It also predicts that there will be a direct relation between the actual frequency of a phrase (the number of times it appears) and processing latencies.

Previous literature on compositional multi-word phrases

There is surprisingly little research on adult processing that can allow us to distinguish between these three approaches. Many studies have shown that two-word (bigram) frequency affects processing. Pronunciation of words is phonetically reduced when the word appears as part of a frequent bigram (Bell et al., 2003, 2009; Gregory et al., 2004; Jurafsky et al., 2001). Also, object relative clauses are processed faster when the embedded clause consists of a frequent subject-verb combination (Reali and Christiansen, 2007). These studies show that people keep track of co-occurrence patterns between single words. But capturing such relations does not require any representation beyond the single word.

Few studies have looked beyond the bigram level. In a seminal study, Bybee and Scheibman (1999) found that *don't* was phonetically reduced in frequently recurring phrases (e.g., *I don't know*). They argued that this provides evidence that very frequent phrases are represented in the lexicon. But although they extracted three-word sequences, they examined the effects of the preceding and following word separately, hence limiting their results to bigrams, which could be modeled without representing larger units. Levy and Jaeger (2007) found that speakers were less likely to produce the optional relativizer in English relative clauses like *How big is the family (that) you cook for?* when the subject of the relative clause (*you*) was more predictable given the previous two words (*the family*). They show that a model that includes the last one, two and three words of the pre-relative clause utterance predicts speakers' use of the optional relativizer, but because they do not report the independent effect of each string size (this was not the goal of their paper), we cannot know whether their results show an effect of three-word

frequency when bigram and unigram frequency are controlled for. Bell et al. (2003) found that words were phonetically reduced when they were more predictable given both the previous and the following word (e.g. in the trigram *middle of the*, the predictability of *of* following *middle* and preceding *the*), again suggesting that speakers represent the expression. But they did not find any effect when looking separately at the predictability a word given the two preceding or two following words. Moreover, this investigation was limited to the ten most frequent words in English.

Underwood et al. (2004) used eye-tracking to look at participants' eye-movements while reading formulaic sequences of up to six words. They compared fixation times for the same word in a formulaic sequence and in a non-formulaic one (e.g., *fact* in: *as a matter of fact* and *it's a well-known fact*). They found fewer fixations when words appeared in formulaic sequences. They interpreted this finding as evidence that people represent the sequences as a whole. But there is an important limitation to this study: the authors did not control for the frequency of the substrings or for the frequency of the bigram that the words appeared in. Since those differed between the formulaic and non-formulaic sequences (e.g., *of fact* and *well-known fact*), the effect could have been driven by bigram frequency rather than phrase frequency.

The only study to control for substring frequency is the one conducted by Bannard & Matthews (2008). Their findings pose a challenge for words-and-rules models: children showed frequency effects for compositional multi-word phrases. But the results are limited in several ways. First, the findings are limited to children. We do not know if the same effects will be found with adults. More importantly, high frequency phrases in this study were taken from the top third of the frequency scale and low

frequency ones from the bottom third. Moreover, only 12 phrases were tested. As they stand, the results do not distinguish between a threshold model and a continuous one - they could still be accommodated if only very frequent phrases were stored.

The effects reported so far provide limited evidence for an effect of phrase-frequency on adults; we need more evidence from adults when substring frequency is controlled for. Moreover, they provide no evidence to distinguish between a threshold model and a continuous one. To do that we need to look at the cases where the two accounts make different predictions: whether very frequent phrases are represented differently from lower frequency ones, and whether frequency of occurrence predicts processing latencies. A continuous model would be preferred if (1) frequency effects were found whenever a higher frequency phrase is compared to a lower frequency one, and (2) there was a clear relation between the actual frequency value and processing latencies.

The current study has several goals. The first is to see if adults are sensitive to the frequency of compositional four-word phrases when the frequency of the smaller parts is controlled for. Such effects are expected under a continuous model. They are also expected if processing reflects expectations derived from linguistic units of varying grain-sizes, including multi-word phrases. The second goal is to distinguish between a threshold model and a continuous one by looking for frequency effects along the continuum (also for lower frequency phrases) and by testing whether actual frequency predicts reaction times across the entire phrase-frequency range. The third goal is a methodological one. Though the effect of frequency on processing is often assumed to be continuous (e.g., Bybee, 2006), in practice, items are often binned into two categories,

high frequency vs. low frequency. By comparing how well a binary measure of frequency (high vs. low) predicts processing latencies compared to a continuous measure, we can test whether the assumption that effects of frequency are continuous is actually supported by empirical RT data, and how much better a continuous measure captures latency differences compared to a binary one.

We investigate these questions by conducting two reaction time experiments where we compare processing latencies for pairs of compositional four-word expressions that differ in phrase frequency (the frequency of the four-word phrase) but are matched for substring frequency (e.g. *don't have to worry* vs. *don't have to wait*). We then conduct a meta-analysis of the reaction times taken from the two experiments to ask whether a continuous measure of frequency predicts processing latencies and whether it does so better than a binary measure.

Experiment 1

We start our investigation by comparing reaction times to multi-word phrases that differ in phrase frequency but are matched for substring frequency. We want to know if people respond faster to higher frequency phrases and if this happens also when comparing phrases in the lower frequency ranges. We look at two frequency bins: in the first bin we set the cutoff point between high and low at ten per million. This is often considered a threshold for representation in the lexical bundle literature (Biber et al. 1999). In the second bin we look at phrases on the lower end of the continuum: in that bin, we set the cutoff point between high and low at one per million. If comprehenders are sensitive to phrase-frequency, they should respond faster to higher frequency phrases. If they store

such information for phrases across the continuum (and not just for very frequent ones) we should see similar effects in the two frequency bins.

We measured processing latencies using a phrasal decision task – people saw four-word expressions and had to judge whether they were possible in English. We chose this task because lexical decision tasks are often used in the study of regularly inflected words (Alegre & Gordon, 1999a; Alegre & Gordon, 1999b; Baayen et al. 1997). Since we are using a similar frequency manipulation (manipulating whole-form frequency) we wanted to also use a similar task. We controlled for the frequency of the substrings by comparing phrases that differed only on the final word and by controlling for the final word, the bigram, and the trigram both in the item selection and in the statistical analysis of the results.

Method

Participants. Twenty-six students (mean age 20 years) from Stanford University participated in the study. All were native English speakers and were paid \$10 in return for their participation.

Materials

We constructed 28 items (16 in the high cutoff bin and 12 in the low cutoff bin). Following Bannard & Matthews (2008), each item consisted of two four-word phrases (we counted orthographic words) that differed only in the final word (*don't have to worry* vs. *don't have to wait*). In each pair the phrases differed in phrase frequency (high vs. low) but were matched for substring frequency (word, bigram, and trigram): the phrases

did not differ in the frequency of the final word, bigram or trigram. Any effect of phrase frequency could not be attributed to a difference in substring frequency. All phrases were constituents of the same kind (two verb-phrases, two noun-phrases, etc) that could form an intonational phrase.

The items were constructed using the Switchboard (Godfrey et al., 1992) and Fisher (Cieri et al., 2004) corpora that were combined to yield a 20 million word corpus. Both corpora consist of American English collected in telephone conversations. We used these corpora because we wanted to create items that were natural (appeared in spontaneous speech), that could form an intonational phrase, and whose frequency was not driven by a specific and uncommon genre (e.g. Wall Street Journal). We selected all the 4-grams that fulfilled the following criteria: (1) the first 3-gram (e.g. *don't have to* had to have a high frequency (over 30 per million); (2) the last word in the 4-gram (the one that differed between the high and low frequency variants) had a frequency of at least 50 per million. Both criteria were used to increase the reliability of the frequency estimates for low frequency phrases, and (3) the last word in the 4-gram could not be a determiner (which would create an incomplete intonational phrase) or a demonstrative. In final position a demonstrative like *that* could also be interpreted as a modifier (e.g. part of *that boy*). Since we do not know what additional processing that may entail, we excluded all such items.

We selected our actual target items by choosing item pairs that had the same first tri-gram, and were matched for the frequency of the final word, bigram, and trigram (see Appendix for complete item list). We used the same corpus to calculate the frequency of the final word, the final bigram and the final trigram of each selected 4-gram. We

selected 16 pairs of four-word phrases for the high phrase-frequency bin, and 12 pairs for the lower phrase-frequency bin. Table 1 shows example items for each bin.

Table 1: Mean frequency (per million words) and example items in the two bins

High bin (High: 19.48, Low: 3.61)		Low bin (High: 3.5, Low: 0.2)	
Don't have to worry	15.3	Don't have any money	2.35
Don't have to wait	1.5	Don't have any place	0.2
I don't know why	35.5	I want to sit	3.6
I don't know who	7.0	I want to say	0.2

High cutoff bin. In this bin, the cutoff point between high and low was set at 10 per million words. In each pair, the high frequency variant appeared over 10 times per million and the low frequency one appeared under 10 times per million. The mean frequency of high frequency phrases in that bin (19.48 per million) was higher than that of low frequency phrases (3.61 per million), $t(30) = 5.86, p < .001$. But there was no difference in the frequency of the final word (high: 922 per million, low: 2235 per million, $t(30) = -1.12, p = .27$), the final bigram (high: 295, low: 174, $t(30) = .73, p = .47$), or the final trigram, (high: 40, low: 23, $t(30) = 1.4, p = .17$) between the high and low frequency phrases. There was also no difference in the number of letters between the high phrases (mean 12.63 letters) and the low ones (mean 12.44 letters), $t(30) = .25, p = .8$.

Low cutoff bin. The cutoff point for the lower-range bin was once per million. We selected 12 pairs of four-word phrases for this bin. In each pair the higher frequency

variant appeared between once and five times per million and the lower frequency variant appeared under once per million. We added an additional restriction that the lower frequency variants had to appear at least twice in the entire 20-million word corpus (.1 per million words). The mean frequency of high frequency phrases (3.5) was higher in that bin than the mean frequency of the low frequency ones (0.2), $t(22) = 4.66, p < .001$. There was no difference in the frequency of the final word (high: 433 per million, low: 278 per million, $t(22) = 1.08, p = .29$), or the final bigram (high: 76, low: 32, $t(22) = 1.44, p = .14$) between the high and low frequency phrases. There was also no difference in the number of letters (high: 12.75 low: 12.33), $t(22) = .44, p = .66$.

Plausibility. We also wanted to control for the real-world plausibility of the events depicted by the low and high frequency phrases. To do so, we used an online survey to gather plausibility ratings for all the selected items. 25 different participants rated the selected items for plausibility on a scale from 1 – 7 (1 - highly implausible, 7 – highly plausible). Plausibility was defined as “describing an entity or situation that is likely to occur in the real world”. Selected items had a high plausibility rating in both bins (high cutoff bin: 6.66, low cutoff bin: 6.51). Importantly, high and low frequency phrases were rated as equally plausible in the high cutoff bin (high: 6.7, low: 6.7, $W = 113.5, p > .5$), and in the low cutoff one (high: 6.6, low: 6.4, $W = 43.5, p > 0.1$, we used Wilcoxon tests because the ratings were skewed towards the plausible end of the scale).

Fillers. In addition to the 56 target items (16 pairs in the high cutoff bin, 12 in the low one, and two variants per pair) we created 80 four-word fillers. Twelve of them were

possible phrases in English and 68 were impossible ones. This design resulted in an equal number of possible and impossible sequences. Fillers were impossible either because they had scrambled word order (e.g. *I saw man the*) or because they had inappropriate prepositions (e.g. *jump during the pool* or *put from the shelf*). Of the impossible fillers, 75% were scrambled and 25% had a wrong preposition. Fillers were chosen so as not to overlap lexically with the experimental items.

Procedure

The experiment was run using Linger (developed by Douglas Rhode, <http://tedlab.mit.edu/~dr/Linger>). Participants sat in a quiet room in front of a computer and completed a phrasal-decision task. In this task participants saw four-word phrases on the screen (in centered position) and had to decide (as quickly as possible) whether they were possible sequences in English. Phrases appeared in their entirety (all four words appeared at once on the screen) and response times were measured from the moment the phrase appeared. Participants used a keyboard to indicate if the phrase was a possible sequence or an impossible one. They used the ‘j’ key to make a ‘yes’ response and the ‘f’ key to make a ‘no’ response (the keys were equally positioned). *I saw the man* was given as an example of a possible sequence. *I saw man the* and *jump during the pool* were given as impossible examples. It was stressed that the sequences did not have to be full sentences to be judged as possible. Participants saw six practice items at the beginning of the experiment.

Each participant saw all 56 experimental items (the 28 item pairs). The task was divided into two blocks with one variant of each item appearing in each block. This was

done to reduce possible priming effects from seeing two very similar phrases (e.g. don't have to worry and don't have to wait). The order of presentation of the blocks was counter-balanced between participants (half saw Block 1 first and half saw Block 2 first). Each block took about 5 minutes to complete. The blocks were separated by an unrelated lexical decision task which took five minutes to complete. None of the words from the main experiment were used in the distracter task. None of the words in the lexical decision task were semantically or phonologically related to the final words in the target phrases.

Results

Responses under 200 ms and over 2 standard deviations from the mean per condition (high vs. low) were excluded. This resulted in the loss of 5% of the data. Accuracy for target items was at ceiling in both frequency bins for high (99%) and low (99%) items. The results are presented separately for the two frequency bins. We analyze the results using mixed-effect linear regression models with subject and item as random effects. We used log (response times) as the predicted variable to reduce the skewness in the distribution of response times. We added the log frequency of the final word and the final bigram as controls (we used log frequencies to correct for non-normal distributions). These were the only substrings that differed between the high and low variant of each pair, and while their frequency was matched, we wanted make sure that any effect of frequency was caused by phrase-frequency and not substring-frequency. These substring-frequencies were calculated from the same corpus used to select the target items.

In all analyses, we checked for collinearity between the fixed effects (for example between phrase-frequency and the frequency of the final bigram), and reduced it by regressing one of the collinear factors (the factor of interest, if one was involved) against the collinear covariates, and using the residuals of these regressions instead of the original variables in the final models we report (this reduced all correlations between factors to less than .19) In all analyses, we also tested whether adding a random interaction (slope) between frequency and subject improved the model (Quene and van de Bergh 2008). The random slope did not significantly improve the model that included all factors of interest and control factors (by model comparison) in any of the analyses and was consequently excluded.

High-cutoff bin. As predicted, participants were sensitive to phrase-frequency. They were faster to respond to high frequency phrases (mean 1040 ms) than to lower frequency ones (mean 1100 ms). The final model had phrase-frequency (high vs. low), log(final-bigram), log(final-unigram), number-of-letters, and block-order (whether the item was seen in the first or the second block of the experiment) as fixed effects, and subject and item as random effects. Because low frequency was coded as the baseline, we expect the coefficient to be negative, indicating that high frequency trials were faster (took less time). The model showed a significant effect of phrase-frequency: participants were faster on high frequency phrases when controlling for substring frequency, number of letters, and order-of-presentation, $\beta = -.053$ (SE = .02), $p < .05$. Model comparisons using the likelihood ratio test showed phrase-frequency to be a significant predictor; the full

model with phrase-frequency fit the data better than a model without it, $\chi^2(1) = 8.93$, $p = .002$.

In addition to phrase-frequency, response times were affected by block order, and number of letters. Decision times were faster in the second block, $\beta = -0.11$ (SE = .02), $p < .001$. Importantly, block-order did not interact with phrase-frequency ($p = .65$); phrase-frequency affected both blocks similarly (we also conducted analyses on each block separately, and all reported results still held). Unsurprisingly, decision times were slower for phrases that had more letters, $\beta = .03$ (SE = .01), $p < .001$, with slower decision times for longer phrases. The frequencies of the final word and the final bigram were highly correlated, $r = .42$, $p = .02$. As a result, the coefficient estimates are not necessarily reliable. Instead we report the results of model comparisons using the likelihood ratio test (comparing the full model to one without the final word and one without the final bigram) which show that neither the final word, $\chi^2(1) = 2.11$, $p = .14$, nor the final bigram, $\chi^2(1) = .004$, $p = .94$, were significant predictors of processing latencies.

Low cutoff bin. As predicted by the continuous approach, but not by the threshold model, participants were also sensitive to phrase-frequency in the lower frequency range. They responded faster to items of higher (but still low) frequency (mean 1059 ms) than lower frequency (mean 1125).

As in the previous bin, we wanted to control for the frequency of the final word and bigram (those differ between the high and low variant in each pair). We ran a mixed-effect model with phrase-frequency (high vs. low), $\log(\text{final-bigram})$, $\log(\text{final-unigram})$, number-of-letters, and block-order (whether the item was seen in the first or the second

block of the experiment) as fixed effects, and subject and item as random effects. We took the same measures as in the previous analysis to reduce any collinearity between the fixed effects.

In this bin also, participants were faster to respond to phrases of higher frequency, $\beta = -.06$ (SE = .02), $p < .02$. Model comparisons using the likelihood ratio test showed phrase frequency to be a significant predictor, $\chi^2(1) = 8.09$, $p = .004$. Again, a negative coefficient is expected because that shows that higher frequency phrases have shorter reaction times.

As in the previous bin, response times were also affected by block order, and number of letters. Decision times were faster in the second block, $\beta = -0.07$ (SE = .02), $p < .001$, and block-order did not interact with phrase-frequency ($p = .51$; we also conducted analyses on each block separately, and all reported results still held). Decision times were slower for longer phrases, $\beta = .03$ (SE = .01), $p < .001$). Because they were correlated, $r = .67$, $p < .001$, we estimated the effect of final word and final bigram using model comparisons using the likelihood ratio test (by comparing a full model to one without the final word and to one without the final bigram). Both final word, $\chi^2(1) = .16$, $p = .68$, and final bigram, $\chi^2(1) = .73$, $p = .39$ were not significant predictors.

Discussion

Experiment 1 set out to test the predictions that people are sensitive to phrase-frequency, and that this is true not only for ‘special’ very-frequent phrases, but for phrases across the frequency continuum. The results showed an effect of phrase-frequency on recognition times for phrases of varying frequency. Since substring frequency was controlled for, the

effect could not have been driven by the frequency of the substrings the phrase is made up of. Since the high and low frequency phrases were also matched for real-world plausibility, it is unlikely that responses reflected knowledge about the frequency of the events depicted by the phrases. These are the first findings to show that four-word phrase-frequency affects adult processing latencies. They mirror effects found for children (Bannard & Matthews, 2008), suggesting that sensitivity to phrase-frequency is not limited to the developing lexicon. They add multi-word phrases to the units that influence processing latencies.

These results provide evidence against a words-and-rules model of representation (Pinker, 1999) – frequency effects were found for linguistic units that can be easily generated from their parts. Finding that more frequent phrases on the lower frequency range were also responded to faster is not compatible with a threshold model where only linguistic units of sufficient frequency can attain independent representation (Biber et al., 1999; Goldberg, 2006; Wray, 2002). Instead, the results are more compatible with a continuous model, where multi-word phrases are one of the many linguistic patterns that are learned and represented.

The results of Experiment 1 showed that language users are sensitive to phrase-frequency on the high and low end of the frequency scale. In Experiment 2 we look at phrases from the middle of the frequency range that fall between the frequency ranges tested in Experiment 1. If phrase-frequency effects are found along the continuum, as predicted by the continuous approach, then mid-frequency phrases should be recognized faster than lower frequency ones. Experiment 2 serves an additional goal. We want to test the prediction that actual phrase-frequency will predict representation strength. To do this,

we need observations for phrases across the frequency continuum. Looking at mid-frequency phrases will complement the low and high frequency phrases tested in Experiment 1. By gathering observations for item across the frequency continuum, we will also be able to conduct a methodological investigation of the relative merit of using a continuous measure of frequency (as opposed to a binary one) in predicting processing latencies.

Experiment 2

In this experiment we looked at the effect of phrase-frequency for a third frequency bin in between the high and low ranges of the first experiment. We set the cutoff between high and low frequency items at five per million (in Experiment 1 the cutoff points were one per million for the lower bin, and ten per million for the higher one). High frequency phrases appeared between five and ten times per million, and low frequency ones appeared between once and five times per million.

Participants

Twenty-three students from Stanford University participated in the study. All were native English speakers and were paid \$10 in return for their participation.

Materials

We constructed 17 target items (see Appendix B for full list). As in Experiment 1, each item consisted of two four-word phrases that differed only on the final word. In each pair, the phrases differed in phrase-frequency but did not differ in frequency of the final word,

bigram, or trigram, The items were constructed using the same corpus and the same selection criteria used in Experiment 1. Only the cutoff point distinguishing high and low frequency items was changed (to 5 per million). High frequency phrases appeared between five and ten times per million. Their low frequency counterparts appeared under five times per million.

The mean frequency of high frequency phrases (7.6) was higher than the mean frequency of the low frequency phrases (2.0), $t(32) = 12.24$, $p < .001$. There was no difference in the frequency of the final word (high: 2445 per million, low: 2267 per million, $t(32) = .25$, $p = .8$), the final bigram (high: 658, low: 424, $t(32) = 1.05$, $p = .3$), or the final trigram, (high: 44, low: 26, $t(32) = .96$, $p = .34$) between the high and low variants. There was also no difference in the number of letters (high: 12.94, low: 13.06), $t(32) = -.16$, $p = .86$. As in Experiment 1, the selected items were also matched for real-world plausibility. 25 different participants rated the selected items using the same online survey used in Experiment 1. Selected items had a high plausibility rating (mean 6.05). Importantly, high and low frequency phrases were rated as equally plausible (high: 6.5, low: 6.3, $W = 101.5$, $p > .1$). We used the same fillers as in Experiment 1.

Procedure

The procedure was identical to Experiment 1. Participants completed a phrasal-decision task in two blocks. One variant of each item appeared in each block. Unlike the previous study, the blocks were separated by a Stroop task that took 5 minutes to complete. In the Stroop task, participants have to give the font color of color words that appear on the screen. In Experiment 1 there was a strong effect of block-order: participants were much

faster in the second block. The effect of block-order did not interact with the effect of frequency but we wanted to see if it would be reduced if we changed the distracter task from a lexical-decision task to a task that did not involve explicit linguistic judgment, like the Stroop task (however, block-order effects were not significantly reduced).

Results

Responses under 200 ms and over 2 standard deviations from the mean per condition (high vs. low) were excluded. This resulted in the loss of 5% of the data. Accuracy for target items was at ceiling for high (99%) and low (98%) items. As in Experiment 1 we analyzed the results using mixed-effect linear regression models to predict logged reaction-times.

The results showed a similar pattern to that of Experiment 1. Participants were faster to respond to higher frequency phrases (1198 ms) compared to lower frequency ones (1276 ms). As in the previous experiment, we wanted to control for the frequency of the final word and bigram (those differ between the high and low variant in each pair). We ran a mixed-effect model with phrase-frequency (high vs. low), $\log(\text{final-bigram})$, $\log(\text{final-unigram})$, number-of-letters, and block-order (whether the item was seen in the first or the second block of the experiment) as fixed effects, and subject and item as random effects.

In this experiment also, participants were faster to respond to phrases of higher frequency, $\beta = -.053$ (SE = .02), $p < .01$. Model comparisons using the likelihood ratio test showed phrase frequency to be a significant predictor, $\chi^2(1) = 9.25$, $p < .01$. The negative coefficient shows that indeed, high frequency phrases had faster reaction times.

As in the previous experiment, response times were affected by block order. Decision times still were faster in the second block despite changing the intervening task, $\beta = -0.08$ (SE = .01), $p < .001$, and block-order did not interact with phrase-frequency ($p = .81$; we also conducted analyses on each block separately, and all reported results still held). Number-of-letters was not a significant predictor, $\beta = .005$ (SE = .007), $p = 0.5$. As in Experiment 1, the frequency of the final word and the final bigram were highly correlated, $r = .42$, $p = .01$. We estimated the effect of the final word and final bigram using model comparisons using the likelihood ratio test (by comparing a full model to one without the final word and to one without the final bigram). The effect of the final word was significant, $\chi^2(1) = 5.9$, $p = .01$, but the effect of the final bigram was not, $\chi^2(1) = 1.17$, $p = .27$.

Discussion

As in Experiment 1, phrase-frequency had a significant effect on reaction times: participants were faster to respond to mid frequency phrases than to lower frequency ones. Whether the cutoff point between high and low was set at ten, five (in Experiment 1), or one per million (in Experiment 2), participants were faster on higher frequency phrases. Experiment 2 provides additional evidence that people store information about compositional multi-word phrases across the frequency range, and that their frequency influences processing.

We now have responses for phrases along the frequency continuum; from ones appearing less than once per million to ones appearing over ten times per million. Table 2

shows the ranges and the means of the three frequency bins we tested in Experiments 1 and 2.

Table 2: Item properties in the three frequency bins (in words per million)

Frequency bin	Condition	Median	Mean	Range
Lo	Hi	2.5	3.5	1.3-9.7
	Lo	0.2	0.2	0.1-0.4
mid	Hi	7.1	7.6	5.4-9.8
	Lo	1.9	2.0	0.8-4.1
hi	Hi	15.4	19.5	9.1-44.8
	Lo	2.8	3.6	0.6-8.9

We can use these data to test a prediction put forth by specific usage-based models that actual frequency predicts representation strength (Bybee, 2006). If true, then the higher the frequency, the faster recognition times should be. We did not test this prediction in Experiments 1 and 2. While we looked at phrases along the frequency continuum, we only used a binary measure to model the results. In each bin we compared high frequency to low frequency. To test the prediction that actual frequency of occurrence predicts reaction times, we conducted a meta-analysis the items from Experiments 1 and 2, using $\log(\text{frequency})$ as a predictor. This investigation serves another goal. The effect of frequency on language processing is often assumed to be continuous. But it is often modeled using binary measures (e.g., Grainger, 1980; Schilling, Rayner & Chumbley, 1998, and many more). We can now test whether the assumption that effects of frequency are continuous is actually supported by empirical reaction time data, and how much better it captures differences in processing latencies compared to binary groupings

Meta-analysis

We performed a meta-analysis of the items used in Experiments 1 and 2. By taking the items from both experiments we have observations for phrases across the frequency range. We now have a more flat distribution of frequencies (with items spread out equally along the frequency continuum) instead of a bimodal one (with items divided by an arbitrary cutoff point). This allowed us to test 1) whether actual frequency predicted reaction times and 2) if it was a better predictor than a binary frequency measure.

Data and Materials

We used all the items from Experiments 1 and 2 in the meta-analysis. This yielded 45 pairs of 4-word phrases (each pair had the same first 3 words). We took the reaction times of all the participants from Experiments 1 and 2 (49 native English speakers). We only used the trials that were included in the analyses of the previous experiments (excluding trials that were answered incorrectly and ones with reaction times above 2 standard deviations from the mean). This yielded a total of 2105 trials.

Results

We analyzed the data using mixed-effect linear regression models, as in Experiments 1 and 2. As a first step we wanted to see whether log (phrase-frequency) was a significant predictor of reaction times. As a second step, we wanted to see if it was a better predictor than a binary measure. To do this, we conducted a breakpoint regression analysis to find the breakpoint that best fits the data (the one where two models fit on either side of it have the maximum summed likelihood, Baayen, 2008). We then compared how well that

binary measure (set at 4.94 per million) fared in comparison with a continuous measure (log phrase-frequency). By choosing the binary measure that is based on the most likely breakpoint in the data, we are biasing against our prediction that a continuous measure will be a better predictor of processing latencies. Before conducting these analyses we had to address potential confounds arising from the use of items that are no longer as well controlled (we are no longer comparing pairs that differ only in phrase-frequency).

Reducing collinearity and over-fitting. In the regression models used in Experiments 1 and 2, we controlled for the frequencies of all the substrings that differed between the low and high variants (they only differed on the final word and bigram). Now that we are treating frequency as a continuous variable, each item pair is effectively treated as two items and the differences between items taken from different pairs are much greater. Items now differ also on the first tri-gram (e.g. *don't have to worry* vs. *go back to work*). To look at the role of phrase-frequency we need to control for the frequencies of all the smaller elements (the two tri-grams, the three bigrams, the four unigrams). But this would risk over-fitting the results (we now have nine frequency measures in addition to the continuous phrase-frequency measure, the binary frequency measure, and the number-of-letter and block-order variables).

To address this, we ran a model with all the variables except the two phrase-frequency ones (the nine frequency controls, block-order, and number-of-letters) as fixed factors and with log (reaction-time) as the dependent variable. Following Baayen (2008), we then removed all the variables whose standard error was greater than the value of their coefficient in the model. This left us with six variables: four frequency control variables

($\log(\text{unigram3})$, $\log(\text{unigram4})$, $\log(\text{bigram1})$, and $\log(\text{trigram1})$), block-order, and number-of-letters. These six variables will eventually go into the full model that will include the two phrase-frequency variables (continuous and binary), along with the random effects of subject and item. In addition to block-order and number-of-letters, some of the frequency control factors still significantly (or marginally) predicted reaction times, even when 4-gram frequency was included: $\log(\text{unigram4})$, $p < .05$ and $\log(\text{trigram1})$, $p < .09$.

Analyzing the results. As predicted by usage-based approaches, continuous log phrase-frequency of occurrence was a significant predictor of reaction times. We established this by comparing a full model with $\log(\text{phrase-frequency})$ and the control variables (not including the binary measure for now), to a model without the continuous frequency measure. The difference between the models was significant, $\chi^2(1) = 14.86$, $p < .001$. Figure 1 plots the model fit for the reaction times to all phrases (note that the mean reaction times for each bin are not corrected for the effect of all control variables, which is probably why they don't form a clean linear trend).

Figure 1: Model fit for reaction times to all phrases

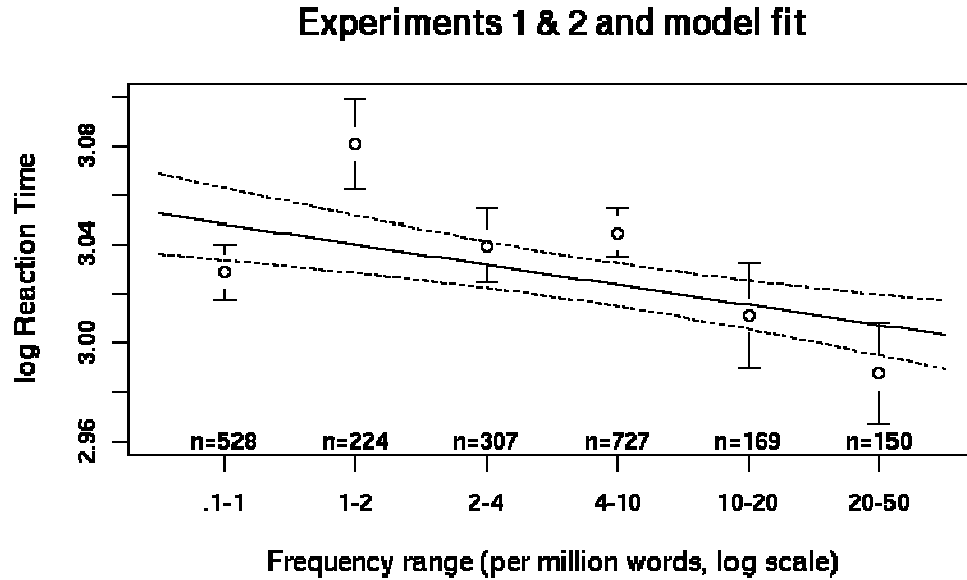


Figure 1: Log reaction time by sequence frequency bin (log scale). Circles represent the means for each bin, with 95% confidence intervals. The fit line is derived from a regression model with a continuous measure of frequency and all control covariates, and also includes 95% confidence intervals.

We now wanted to see whether the continuous variable accounted for more of the variance than the binary one. To do this we compared a full model with both frequency measures (and the control variables) once to a model without the continuous measure, and once to a model without the binary measure. The likelihood ratio comparisons showed that the continuous frequency was a significant predictor, $\chi^2(1) = 10.5, p < .01$, but the binary measure was not, $\chi^2(1) = .69, p > .4$. In other words, when continuous frequency was taken into account, the binary measure was no longer significant; it didn't explain any additional variance.

Discussion

These results are in accordance with a usage-based approach where every additional occurrence of a sequence strengthens its activation. The results also show that a continuous measure of frequency is a better predictor of reaction times than a categorical one, even when analyzed with a conservative model that controls for the frequencies of all the substrings and words making up the four-word sequence, as well as participant and item effects. They provide qualitative support for the commonly accepted observation that the effect of frequency on language processing is continuous. This is not a controversial idea but there have been no empirical investigations that we know of that pit a binary variable against a continuous one. It also reinforces statistical argumentation against dichotomizing continuous variables (e.g. Cohen and Cohen, 1983). Methodologically, the results highlight the advantage of (1) using continuous measures of frequency as predictors of reaction times, and (2) using statistical tools of analysis (e.g., regression models) where such continuous predictors can be used.

General Discussion

We set out to investigate whether the frequency of compositional four-word expressions affects processing, and whether these effects are found not only for ‘special’, very frequent phrases, but whenever a higher frequency phrase is compared to a lower one. Experiments 1 and 2 and the meta-analysis provided an affirmative answer to both questions: higher frequency phrases were responded to faster. The effect was found across the frequency range and was a gradient one. It was better captured when frequency was not binned, but treated as a continuous variable. The meta-analysis showed a direct

relation between frequency of occurrence and processing latencies: the more often a phrase has been experienced, the faster it was processed.

These effects cannot be attributed to substring frequency – the pairs of phrases we compared were matched for the frequency of all substrings. It is also unlikely that they reflect a difference in the real-world likelihood of the events depicted by the phrases since they were rated as equally likely/plausible. Because the phrases differed on the final word, it would not be enough to store co-occurrence information for words or even two-word sequences. To recognize that one phrase is more frequent than the other, one would need to know how often the entire phrase appears. That is, store co-occurrence information for at least four consecutive words.

These results advance our understanding of frequency effects in several ways. First, they show frequency effects for four-word phrases. Though emergentist models (e.g. usage-based, connectionist, exemplar) have been around for almost 20 years, there has been little empirical research testing their predictions for larger chunks of language. Our findings fill an empirical gap, since phrase-frequency effects are predicted under such models, but were not previously reported. Second, our results highlight the loss of power when frequency is treated as a binary variable. While frequency is often viewed as a continuous variable, in practice, items are generally binned into two categories, high frequency vs. low frequency. By pitting a binary measure against a continuous one we demonstrate the advantage of using a continuous measure as a predictor of processing latencies.

What information are language users sensitive to?

At a minimum, the current findings add multi-word phrases to the units that influence processing in adults. They show that language users are sensitive to co-occurrence patterns beyond the bigram level. This raises questions about how to integrate different frequency measures in a processing model, and how to capture and predict phrase-frequency effects when modeling linguistic knowledge.

As they stand, the results cannot be taken as evidence that the phrases were accessed as unanalyzed wholes – we do not know whether, and to what extent, the parts of the phrases were activated. Our experiment was not designed to test this - our experimental items were designed to keep substring frequency maximally similar within an item pair (e.g. *don't have to worry* vs. *don't have to wait*). Looking back at the results of the meta-analysis, there is evidence that substring frequency (the frequency of the fourth word and the first trigram) still affected reaction times when controlling for phrase-frequency, as would be expected given the wide-spread word frequency effects in reading where words are read as part of a larger linguistic context. Nor can we know whether the phrases were processed incrementally faster due to their increased predictability: we only obtained reaction times for the entire phrase. More work using different experimental paradigms such as self-paced reading is needed to study the way part and whole frequency interact and the way phrase-frequency effects arise over time.

Word frequency effects led to revised models of lexical access (Gaskell & Marslen-Wilson, 1997; McClelland & Elman, 1986). Finding that speakers maintained information about the voice a word was uttered in led to the creation of episodic models of the lexicon (Goldinger, 1996). Finding speaker-dependent phonetic effects fueled the

development of phonetic exemplar models where such variability can be accommodated (Pierrehumbert, 2001). Finding phrase-frequency effects can have a similar effect of extending existing models. Many of the models currently available focus on modeling frequency effects at the word level or below, or at the level of syntactic constructions. To capture phrase-frequency effects, such models would have to incorporate larger frequency relations.

One possibility, in line with exemplar models of language (Bod, 1998; Goldinger, 1996; Johnson, 1997; Pierrehumbert, 2001, 2006), is to implement the representations produced by the exemplar-based syntactic models of Bod (1998, 2006) in a spreading-activation network, as proposed in Snider (2008). In the model that Bod presents, syntactic productivity is achieved by starting with arbitrarily large linguistic units and deducing syntactic structure by means of statistical inference. The resulting lexicon has structurally analyzed chunks of different grain-sizes, along with a mechanism for constructing larger structures out of them. The processing of units is influenced by the probability of the smaller units used to form them (Bod 2006).

Implementing these representation in a spreading-activation network (Snider, 2008) will result in patterns of varying levels of abstraction (from fully realized strings of words, to fully abstract constructions) that are linked to each other, and whose activation is related to frequency of occurrence. Multi-word phrases are naturally represented in this model, and are linked to the words and smaller strings they consist of, as well as to the more abstract constructions they participate in. The same would apply to all phrases, regardless of their frequency, and would lead to complementary representations at different grain sizes.

In this model, frequency effects on processing reflect a complex interaction between the various frequency measures of the different grain-sizes. This fits in nicely with previous findings that processing latencies are better captured when frequencies at different grain-sizes are taken into account simultaneously. For example, many studies have found that comprehension is affected by the sub-categorization biases of the verb (Trueswell & Tanenhaus, 1994; MacDonald, 1994; Garnsey et al., 1997; Gahl, 2002). But not all studies find such effects (Ferreira & Henderson, 1990; Kennison, 2001; Pickering et al 2000). Hare et al. (2003, 2004) suggest that the empirical discrepancies arise because verb-sense (an additional grain-size) was not taken into account. In similar spirit, Crocker and Brants (2000) reinterpret the early preference for a noun-phrase continuation regardless of verb-bias reported in Pickering et al. (2000) as reflecting the overall lower probability of the S-complement analysis (Hale, 2003 and Jurafsky, 1996 also include the frequency of syntactic rules in their parsing models). The model described above naturally captured frequency estimates on multiple levels of linguistic analysis.

Phrase-frequency effects could also be modeled in other ways. It may be possible to capture four-word frequency effects in a simple recurrent network (Elman, 1991). Such networks have been used successfully to model syntactic processing (e.g., object vs. subject relative clauses, MacDonald & Christiansen, 2002). Combined with a dynamic system approach, they have been used to model local-coherence effects (Tabor et al, 1997) and thematic fit effects (Tabor & Tanenhaus, 1999) both of which require keeping track of the distributional contexts that words appear in. Phrase-frequency effects may be

similarly modeled by keeping track of larger ‘frequency’ chains. However, there are no existing connectionist networks that set out to capture such relations.

Incorporating linguistic units at varying grain sizes raises interesting questions about the relation between multi-word phrases and two kinds of linguistic units (1) the substrings they are made of (e.g. ‘*to worry*’) and (2) the more abstract units they are instances of (e.g. an infinitive clause). Kapatsinski & Radicke (2008) argue for competition between larger units and their parts when the whole-form is of sufficient frequency. Participants had to respond whenever they detected the particle ‘up’ in a verb-particle combination (e.g. give up). Reaction times were faster the more frequent the collocation. But for collocations in the highest frequency bin, there was a slowdown in reaction times. A similar result is reported in a separate study on the function word *of* (Sosa and MacFarlane, 2002): detection of *of* was slower in highly frequent collocations like kind of (ones that appeared over 800 times per million). These findings are interpreted as evidence for competition between the part and the whole when the whole is frequent enough. Reaction times speed up when the particle is more predictable, given the sequence, but slow down when there is competition between the particle and the “chunked” collocation. This study suggests an interesting way to reconcile claims about the ‘special’ status of very frequent units (e.g. Bybee, 2002; Goldberg, 2006) with the current findings. We did not find evidence for a distinction between very frequent phrases and lower frequency ones: phrase-frequency effects were found across the continuum. However, very frequent phrases may differ in the degree to which the parts activate the whole and vice versa.

The representational status of multi-word phrases: evaluating the evidence

We set out to distinguish between three views on the representational status of multi-word phrases. The current findings are hard to accommodate within a words-and-rules model where compositional units (regular words or compositional phrases) are not expected to display whole-form frequency effects. They are not easy to accommodate within a threshold model that posits a unique status for very frequent forms. There was no indication of a clear difference between very frequent phrases and lower frequency ones: frequency effects were found for all the tested phrases. Using a frequency threshold as a determiner of storage is also problematic because speakers cannot know a-priori which phrases will become frequent enough to merit storage. Whatever information is maintained for very frequent phrases must have been once registered for all phrases. This information could be discarded at later stages of learning, but this seems unlikely. A similar argument can be made against using idiosyncrasy of meaning as a criterion for inclusion in the lexicon (Goldberg, 2006; Wray, 2002). From the perspective of the child learner who has yet to hone in on the grammatical regularities of his/her language, all linguistic input starts out being idiosyncratic and ‘irregular’ to some degree. However, our findings do not rule out threshold models where the relation between the parts and the whole changes according to whole-frequency. This is not something we tested in the current paper.

The findings are most compatible with a continuous model of representation where frequency is expected to affect all linguistic forms in a similar way. Compositional phrases showed whole-form frequency effects like those displayed by simple and inflected words. Furthermore, we found no evidence for a dichotomous distinction

between very frequent phrases and all other phrases. More broadly, these findings argue against a clear distinction between the linguistic forms that are ‘stored’ and the ones that are ‘computed’. Instead, they enhance an emergentist view where all linguistic material is represented and processed in a similar fashion.

The distinction between ‘stored’ and ‘computed’ material is further blurred by recent findings on the processing of idioms. Idioms are often seen as prototypical candidates for ‘storage’ (Pinker, 1999; Jackendoff, 1995 but see Nunberg, Sag & Wasow, 1995 for an argument that only few idiomatic phrases are truly non-compositional). However, recent several recent experimental results reveal parallels between the processing of idiomatic and non-idiomatic phrases. Sprenger et al (2006) show that idioms can prime and be primed by words that appear in them (e.g. *hit the road* prime *road*), suggesting that like compositional phrases, they have internal structure. Konopka & Bock (2009) show that idiomatic and non-idiomatic phrasal verbs (e.g. *pull off a robbery*) can prime particle placement (whether the particle appears before or after the direct object) in non-idiomatic phrases that have no lexical overlap (e.g. *knocked over the vase*). Using acceptability judgment of familiar and invented idioms Tabossi, Wolf, & Koterle (2009) argue that the syntax of idioms is governed by the same syntactic and pragmatic principles that govern non-idiomatic language. These findings highlight the difficulty in distinguishing between ‘stored’ and ‘computed’ forms.

The difficulty in finding a clear criterion for inclusion in the lexicon leads Elman (2009) to the radical solution of “lexical knowledge without a lexicon”. Elman reviews numerous studies detailing the rich information language users have about verbs (from the agents it appears with to the discourse situation it evokes), and the way this

information is rapidly used in online processing. The rapid availability of such detailed, situation-specific lexical information suggests that “either the lexicon must be expanded to include factors that do not plausibly seem to belong there; or else virtually all information about word meaning is removed, leaving the lexicon impoverished” (pp. 1). Instead, Elman argues for an emergentist model in which linguistic knowledge is viewed as a constantly changing dynamic system and where the lexicon doesn’t contain fixed units but dynamic patterns. We propose that phrasal frequency effects similarly require a model that transcends traditional notions of the lexicon.

Limitations

The results of the current study are limited in that all the phrases that we used were constituents: verb phrases, noun phrases, prepositional phrases. They possessed some structural consistency. This doesn’t in any way undermine the effect of frequency, but we cannot rule out the possibility that people are only sensitive to the frequency of multi-word sequences that are also constituents. This would pose an interesting challenge for emergentist models of language. The phrases were always presented out-of-context. It is likely that like other linguistic units, the processing of multi-word phrases will also be influenced by expectations formed on the basis of prior linguistic context. In fact, finding that manipulating the linguistic context can affect phrase-processing would provide additional support for treating phrases as units of processing.

The results are also limited in that phrases always differed on the final word and that word was always a content word (e.g., *worry*, *wait*). We do not know if the same effects would hold when the phrases differ in function words, or when the words that they

differ on are not in final position. For example, in a corpus study of word duration, Bell et al. (2009) found that different predictability measures affected the duration of function words and content words: both content and function words were shorter when they were predictable given the following word, but only very frequent function words were sensitive to predictability given the preceding word. We see no theoretical reason to suggest that phrase frequency effects will not hold for non-constituents, or will not hold when a different word is in non-final position, but this will require further investigation. On a more basic level, our results do not tell us *why* certain phrases are more frequent than others. They do not address the multiple linguistic and real-world factors that make certain linguistic configurations more frequent, but they show that whatever the underlying causes, frequency differences influence language use.

Implications for parsing, production and learning

Words have served an important role in parsing and production models. Word frequency influences interpretation: parses reflect the more frequent uses of a word (e.g. the garden-path caused by a sentence like *The old man the bridge*, in which *man* is used as a verb). But what if phrase frequency affects parsing in a similar way? For example, ambiguity resolution may be driven not only by how often a verb appears as a past participle and how likely a noun is to be an Agent, but also by the exact frequencies of the noun-verb combination. Patterns such as this have been observed in the processing of object relative clauses where chunk frequency influenced processing speed (Reali and Christiansen, 2007). If the effect of chunk frequency on parsing is widespread, then (1) parsing models will have to take into account chunk frequency, and (2) chunk frequency may need to be

controlled in experiments. Production models have also assumed that creating an utterance involves a stage of word selection that is separate from the syntactic level (Levelt, 1999). What if multi-word phrases are also selected in production? Speakers' choices could be driven by a tendency to use constructions with higher phrase-frequency. These ideas must for the moment be considered speculative, but the current findings highlight the need to look at the role of multi-word phrases as well as single words in parsing and production.

Words are often highlighted also in acquisition research, as the units that children need to acquire (much research focuses on how children segment words from speech and assign them meaning). Yet multi-word phrases may also play an important role in language learning, especially if grammatical knowledge emerges by abstracting over stored utterances (Abbott-Smith & Tomasello, 2006). Being able to represent and utilize them may assist extracting grammatical regularities (e.g. using frequent frames to learn about grammatical categories, Mintz, 2003), and not doing so may be one of the factors that hinders adult language learning (Arnon & Ramscar, 2009). Finding that multi-word phrases are units of representation for adults thus opens interesting questions about their role in language learning.

Conclusion

This study adds the frequency of multi-word phrases to the distributional information that people have access to during language processing. People responded faster to more frequent four-word phrases at all points across the frequency spectrum: there was no evidence for a threshold beyond which these effects occurred. These findings have

implications for models of processing and representation. They call for processing models that can capture phrase-frequency effects, and highlight the importance of incorporating larger units into parsing and production models. At the same time, they argue for an emergentist model of linguistic knowledge where experience influences the learning, representation, and processing of all linguistic patterns.

Acknowledgment

This work was supported by a Stanford Graduate Fellowship given to both authors, and NSF Award No. IS-0624345. We would like to Dan Jurafsky for guidance and support. We thank Florian Jaeger, Dan Jurafsky, Victor Kuperman Meghan Sumner, and Harry Tily for helpful comments and suggestions, as well as the audience at the 83rd meeting of the Linguistic Society of America.

A. Appendix A

Materials used in Experiment 1 (high frequency range), with the frequency per million words in the Fisher corpus.

1. a lot of places 10.45
a lot of days 0.55
2. a lot of work 14.70
a lot of years 1.90
3. all over the place 21.45
all over the city 0.65
4. don't have to worry 15.30
don't have to wait 1.40
5. don't know how much 12.80
don't know how many 7.80
6. go to the doctor 16.70
go to the beach 5.65
7. how do you feel 29.60
how do you do 4.95
8. I don't know why 35.15
I don't know who 7.00
9. I have a lot 26.45
I have a little 8.95
10. I have to say 15.40
I have to see 0.95

11. I want to go 9.10
I want to know 2.95
12. it's kind of hard 13.30
it's kind of funny 7.20
13. on the other hand 27.15
on the other end 3.95
14. out of the house 9.75
out of the game 0.70
15. we have to talk 9.70
we have to say 0.65
16. where do you live 44.80
where do you work 2.60

Materials used in Experiment 1 (low frequency range), with the frequency per million words in the Fisher corpus.

1. a lot of rain 4.65
a lot of blood 0.20
2. don't have any money 2.35
don't have any place 0.25
3. going to come back 1.35
going to come down 0.40
4. have to be careful 5.90
have to be quiet 0.15

5. I have a sister 4.90
I have a game 0.10
6. I have to pay 1.80
I have to play 0.10
7. I want to say 3.60
I want to sit 0.20
8. it was really funny 2.65
it was really big 0.15
9. out of the car 2.00
out of the box 0.20
10. we have to wait 1.65
we have to leave 0.25
11. we have to talk 9.70
we have to sit 0.20
12. you like to read 1.55
you like to try 0.10

B. Appendix B

Materials used in Experiment 2 (mid frequency range), with the frequency per million words in the Fisher corpus.

1. a lot of problems 9.60
a lot of power 0.85
2. all over the country 9.55

- all over the house 0.75
3. be able to go 8.40
be able to see 3.95
4. do you know how 6.40
do you know when 1.85
5. go back to school 6.75
go back to work 4.00
6. how do you get 6.95
how do you go 1.05
7. I don't really care 6.20
I don't really need 0.85
8. I don't see how 9.20
I don't see them 2.45
9. it takes a lot 7.25
it takes a little 1.45
10. know what that is 6.25
know what that was 1.05
11. not going to get 7.95
not going to see 0.75
12. out of the house 9.75
out of the car 2.00
13. take care of them 6.90
take care of things 0.80

- 14. to have a lot 6.55
to have a little 2.70
- 15. we used to go 7.05
we used to be 2.25
- 16. you know how many 5.40
you know how long 3.40
- 17. you're going to be 9.70
you're going to do 4.05

References

- Abbot-Smith, K. & Tomasello, M. (2006). Exemplar-learning and schematization in a usage based account of syntactic acquisition. *The Linguistic Review*, 23, 275-290.
- Alegre, M. and Gordon, P. (1999a). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language*, 40, 41-61. Elsevier.
- Alegre, M. and Gordon, P. (1999b). Rule-based versus associative processes in derivational morphology. *Brain and Language*, 68, 347-354. Elsevier.
- Arnon, I., & Ramscar, M. (2009). Order-of-acquisition affects what gets learned. In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (pp. xx-xx). Cognitive Science Society.
- Baayen, R.H., T. Dijkstra, and R. Schreuder. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language* 37, 94-117. Elsevier.
- Baayen, H., R. Schreuder, N. deJong, and A. Krott. (2002). Dutch inflection: the rules that prove the exception. In S. Nooteboom, F. Weerman, and F. Wijnen (Eds.), *Storage and computation in the language faculty*. 61–92. Boston: Kluwer
- Baayen, H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.
- Baayen, R., Feldman, L., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, 55, 290-313.

- Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: the effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, 19, 241-248.
- Bell, A., Brenier, J., Gregory, M., Girand, C., Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60, 92-111.
- Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, & M., Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, 113, 1001-1024.
- Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English*. MIT Press.
- Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, 18, 355-387.
- Bod, R., J. and Hay, and S. Jannedy. (2003). *Probabilistic linguistics*. MIT Press.
- Bod, R. (1998). *Beyond Grammar: An Experience-based Theory of Language*. Center for the Study of Language and Information, CA.
- Bod, R. (2001). Sentence memory: Storage vs. computation of frequent sentences. Talk presented at CUNY.
- Bod, R. (2006). An all-subtrees approach to unsupervised parsing. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, 865-872.

- Bresnan, J., Cueni, A., Nikitina, T., & Baayen, H. (2007). Predicting the dative alternation. *Proceedings of the Royal Netherlands Academy of Science Workshop on Foundations of Interpretation*. Amsterdam.
- Bybee, J., and P. Hopper. (2001). *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.
- Bybee, J. (1998). *The emergent lexicon*. In Chicago Linguistic Society, 34, 421-435.
- Bybee, J. (2002). Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition*, 24, 215-221.
- Bybee, J. (2006). From usage to grammar: The minds response to repetition. *Language*, 82, 711-733.
- Bybee, J., & McClelland, J. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review*, 22, 381-410.
- Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: the reduction of don't in English. *Linguistics*, 37, 575-596.
- Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, 113, 2, 234-272.
- Christiansen, M.H. and N. Chater. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23:157-205.
- Cieri, C., Miller, D & Walker, K. (2004). The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text. *In Proceedings of LREC 2004*.
- Clifton, C., Frazier, L., & Connine, C. (1984). Lexical expectations in sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, 23, 696-708.

- Cohen, J. and P. Cohen. (1983). *Applied multiple regression-correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Crocker, M. & Brants, T. (2000). Wide Coverage Probabilistic Sentence Processing. *Journal of Psycholinguistic Research*; 29, 647-669.
- Dell, G. S. (1986). A spreading activation theory of retrieval in language production. *Psychological Review*, 93, 283-321.
- Dell, G. S., Chang, F., and Griffin, Z. M. (1999). Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science*, 23, 517-542.
- Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, 25, 104-123. Elsevier.
- Ellis, N. (2002). Frequency Effects in Language Processing. *Studies in Second Language Acquisition*, 24, 143-188.
- Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, 14, 179-211.
- Elman, J.L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7, 195-225.
- Elman, J.L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33, 547-582.
- Ferreira, F., & Henderson, J.M. (1990). The use of verb information in syntactic parsing: Evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 555-569.
- Frazier, L., & Fodor, J.D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, 66, 291-325..

- Gahl, S. (2002). Lexical biases in aphasic sentence comprehension: An experimental and corpus linguistic study. *Aphasiology*, 16, 1173-1198.
- Gahl, S., and S. M. Garnsey. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language* 80.748 – 775.
- Gahl, S. and A. C. L. Yu (eds.). (2006). Special theme issue: Exemplar-based models in linguistics. *The Linguistic Review*, 23. 213–379.
- Garnsey, S., Pearlmutter, N., Myers, E., and Lotocky, M. (1997). The Contributions of Verb Bias and Plausibility to the Comprehension of Temporarily Ambiguous Sentences. *Journal of Memory and Language*, 37, 58-93.
- Gaskell, M., & Marslen-Wilson, W. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, 12, 613-656.
- Godfrey, J.J., E.C. Holliman, and J. McDaniel. (1992). Switchboard: Telephone speech corpus for research and development. In IEEE ICASSP , 517–520.
- Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: Chicago University Press.
- Goldberg, A. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.
- Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 22, 1166-1183.
- Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (pp. 1–8). Pittsburgh, PA: Carnegie Mellon University.

- Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32, 101–123.
- Hare, M., K. McRae, and J.L. Elman. (2003). Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language*, 48, 281-303.
- Hare, M., K. McRae, and J.L. Elman. (2004). Admitting that admitting verb sense into corpus analyses makes sense. *Language and Cognitive Processes*, 19, 181–224.
- Hay, J., J. Pierrehumbert, and M. Beckman. (2004). Speech perception, well-formedness, and the statistics of the lexicon. *Papers in Laboratory Phonology*, VI, 58-74.
- Jackendoff, R. (2002). *Foundations of language*. Oxford University Press, New York.
- Jaeger, T. F. (2006). *Redundancy and syntactic reduction in spontaneous speech*. Unpublished dissertation, Stanford University, Stanford, CA.
- Jescheniak, J. and W.J.M Levelt. (1994). Word frequency effects in production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 824-843.
- Johnson, K., (1997). Speech perception without speaker normalization. In K. Johnson & Mullenix (Eds.) *Talker Variability in Speech Processing*. San Diego, Academic Press, 145-166.
- Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20, 137–194.
- Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In Bod, R., Hay, J., and Jannedy, S., editors, *Probabilistic Linguistics*. MIT Press.

- Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In: J. Bybee & P. Hopper, P. (Eds.), *Frequency and the Emergence of Linguistic Structure*. John Benjamins, Amsterdam, 229-254.
- Kapatsinski, V., & Radicke, J. (2008). Frequency and the emergence of prefabs: Evidence from monitoring. In: R. L. Corrigan, E. A. Moravcsik, H. Ouali, K. M. Wheatley (Eds.), *Formulaic Language*. John Benjamins.
- Kennison, S. M. (2001). Limitations on the use of verb information during sentence comprehension. *Psychonomic Bulletin & Review*, 8, 132-138.
- Konopka, A. E., & Bock, J. K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cognitive Psychology*, 58, 68-101.
- Langacker, R. (1987). *Foundations of Cognitive Grammar*. Stanford University Press.
- Langacker, R. (1988). A usage-based model. *Topics in Cognitive Linguistics*, 50, 127-163.
- Levelt, W. (1999). Models of word production. *Trends in Cognitive Sciences*, 3, 223-232.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106, 1126-1177.
- Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In: B. Schlokopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems (NIPS)*, 19, MIT Press, Cambridge, MA, 849-856.
- MacDonald, M. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, 9, 157-201.

- MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676-703.
- MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: A comment on Just & Carpenter (1992) and Waters & Caplan (1996). *Psychological Review*, *109*, 35-54.
- MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology*, *49*, 199-227.
- McClelland, J., & Elman, J. (1986). Interactive processes in speech perception: the TRACE model. In: *Computational Models Of Cognition And Perception Series*. MIT Press, Cambridge, MA, 58-121.
- McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modelling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*, 283-312.
- Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91-117.
- Mitchell, D. C., Cuetos, F., Corley, M. M. B., & Brysbaert, M. (1995). Exposure-based models of human parsing: Evidence for the use of coarse-grained (non-lexical) statistical records. *Journal of Psycholinguistic Research*, *24*, 469-488.
- Monsell, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition*, 148-197. Hillsdale, NJ: Erlbaum.
- Morton, J. (1969) The interaction of information in word recognition. *Psychological Review*, *76*, 165-78.

- Nunberg, G., Wasow, T., & Sag, I. A. (1994). Idioms. *Language*, 70, 491-538.
- Pickering, M., M. Traxler, and M. Crocker. (2000). Ambiguity resolution in sentence processing: evidence against likelihood. *Journal of Memory and Language*; 43, 447-475.
- Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (eds.) *Frequency Effects and the Emergence of Lexical Structure*. John Benjamins, Amsterdam. 137-157.
- Pierrehumbert, J. (2006). The next toolkit. *Journal of Phonetics*, 34, 516-530.
- Pinker, S., and Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, 6, 456-463.
- Pinker, S. (1991) Rules of Language. *Science*, 253, 530-535.
- Pinker, S. (1999). *Words and Rules: The Ingredients of Language*. Basic Books.
- Prince, A., & Pinker, S. (1988). Rules and connections in human language. *Trends in Neurosciences*, 11, 195-202.
- Quene', H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, 59, 413-425.
- Rayner, K., & Duffy, S. A. (1988). On-line comprehension processes and eye movements in reading. In M. Daneman , G. E. MacKinnon, & T. G. Waller (Eds.), *Reading research: Advances in theory and practice*, 13–66. New York: Academic Press.
- Reali, F., & Christiansen, M. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*, 57, 1-23.

- Rodriguez, P., Wiles, J., & Elman, J. L. (1999). A recurrent neural network that learns to count. *Connection Science*, 11, 5–40.
- Rumelhart, D., & McClelland, J. (1986). On learning the past tenses of English verbs. In: D. Rumelhart & J. McClelland, J. (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, 216-271.
- Schilling, H. E. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: word frequency effects and individual differences. *Memory and Cognition*, 26, 1270–1281.
- Seidenberg, M. S. (1994). Language and connectionism: The developing interface. *Cognition*, 50, 385-401.
- Snider, N. (2008). *An exemplar model of syntactic priming*. Unpublished dissertation, Stanford University, Stanford, CA.
- Sosa, A., & MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: collocations involving the word of. *Brain and Language*, 83, 227-236.
- Sprenger, S., Levelt W.J.M., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases, *Journal of Memory and Language*, 54, 161–184.
- Stemberger, J. P., & MacWhinney, B. (1988). Are inflected forms stored in the lexicon? In M. Hammond & M. Noonan (Eds.), *Theoretical morphology: Approaches in modern linguistics* (pp. 101–116). San Diego, CA: Academic Press.
- Tabor, W. and Tanenhaus, M. K. (1999) Dynamical Models of Sentence Processing . *Cognitive Science*, 23, 491-515.

- Tabor, W., Juliano, C. and Tanenhaus, M. K. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes*, 12, 211-271.
- Tabossi, P., Wolf, K. & Koterle, S. (2009). Idiom syntax: idiomatic or principled? *Journal of Memory and Language*, 61, 77-96.
- Taft, M. (1979) Recognition of affixed words and the word-frequency effect. *Memory and Cognition*, 7, 263–72.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995) Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632-1634.
- Tily, H., Gahl, S., Arnon, I., Kothari, A., Snider, N., & Bresnan, J. (to appear). Pronunciation reflects syntactic probabilities: Evidence from spontaneous speech. *Language and Cognition*, 2, xx-xx.
- Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.
- Trueswell, J., & Tanenhaus, M. (1994). Toward a lexical framework of constraint-based syntactic ambiguity resolution. In: C. Clifton, L. Frazier, K. R. (Eds.), *Perspectives on Sentence Processing*. Hillsdale, NJ: Erlbaum, 155-179.
- Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic Influences on Parsing: The use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33, 285-318

- Ullman, M. T., Pancheva, R., Love, T., Yee, E., Swinney, D., & Hickok, G. (2005). Neural correlates of lexicon and grammar: Evidence from the production, reading, and judgment of inflection in aphasia. *Brain and Language*, *93*, 185-238.
- Ullman, M. T. (2001). A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews Neuroscience*, *2*, 717-726.
- Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In: N. Schmitt (Ed.) *Formulaic Sequences: Acquisition, Processing, and Use*. John Benjamins, Amsterdam, 153-172.
- Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge University Press.