Article

Language *and* Speech

More than Words: The Effect of Multi-word Frequency and Constituency on Phonetic Duration

Language *and* Speech
56(3) 349–371

DOI: 10.1177/0023830913484891
las.sagepub.com

Inbal Arnon
University of Haifa, Israel

Uriel Cohen Priva
Brown University, USA

Abstract
There is mounting evidence that language users are sensitive to the distributional properties of multi-word sequences. Such findings expand the range of information speakers are sensitive to and call for processing models that can represent larger chains of relations. In the current paper we investigate the effect of multi-word statistics on phonetic duration using a combination of experimental and corpus-based research. We ask (a) if phonetic duration is affected by multi-word frequency in both elicited and spontaneous speech, and (b) if syntactic constituency modulates the effect. We show that phonetic durations are reduced in higher frequency sequences, regardless of constituency: duration is shorter for more frequent sequences within and across syntactic boundaries. The effects are not reducible to the frequency of the individual words or substrings. These findings open up a novel set of questions about the interaction between surface distributions and higher order properties, and the resulting need (or lack thereof) to incorporate higher order properties into processing models.

Keywords
Speech production, phonetic duration, emergentist models, multi-word frequency

1 Introduction

Recent years have seen an increased interest in emergentist models of language where language structure is intrinsically tied to language use. Whether implemented using connectionist networks (McClelland, 2010; McClelland et al., 2010), exemplar-based representations (Beekhuizen et al., 2013; Bod, 2009), dynamic-system theory (Elman, 2009), or discriminative learning mechanisms

Corresponding author:
Inbal Arnon, Psychology Department, University of Haifa, Haifa, Israel.
Email:Inbal.arnon@gmail.com

(Baayen, Milin, Filipovic Durdevic, Hendrix, & Marelli, 2011; Baayen et al., this volume), such models undermine the traditional distinction between words and rules specifically (Chomsky, 1965; Pinker, 1999)—as well as between the lexicon and grammar more generally—and argue instead that all linguistic material is processed and represented by the same cognitive mechanism. Knowing a language, in such approaches, means acquiring a complex system of patterns—ranging from words through multi-word sequences to more abstract constructions (Bybee, 1998; Goldberg, 2006; Sag, 2013). Language processing is affected by the properties of linguistic elements of various sizes and levels of abstraction—regardless of whether they are atomic (e.g., an uninflected word such as *cat*) or not (e.g., a three-word compositional sentence such as *see the cat*). Consequently, speakers are predicted to be as sensitive to the properties of multi-word sequences (e.g., *I don't know*) as they are to those of single words.

Accordingly, there is growing evidence that speakers (children and adults) are sensitive to the properties of multi-word sequences and draw on such information in production, comprehension, and learning (Arnon & Clark, 2011; Arnon & Snider, 2010; Bannard & Matthews, 2008; Ellis, Simpson-Vlach, & Maynard, 2008; Frank & Bod, 2011; Reali & Christiansen, 2007; Siyanova-Chanturia, Conklin, & van Heuven, 2011; Tremblay & Baayen, 2010; Tremblay, Derwing, Libben, & Westbury, 2011). Adult speakers are faster to recognize higher frequency four-word phrases (Arnon & Snider, 2010) and show better memory of them (Tremblay & Baayen, 2010), even when part frequencies are controlled for. Young children (two- and three-year-olds) are faster and more accurate at producing higher frequency phrases compared to lower frequency ones (Bannard & Matthews, 2008), while slightly older children show better production of irregular plurals inside frequent frames (e.g., *Brush your – teeth*, Arnon & Clark, 2011).

Finding that speakers are affected by larger distributional patterns (e.g., the frequency of multi-word strings) has theoretical and empirical implications. From a theoretical perspective, the findings illustrate parallels in the processing of words and larger sequences (Snider & Arnon, 2012) and make it harder to argue, as traditional models do, that the forms are generated and processed by two qualitatively different cognitive systems (Pinker & Ullman, 2002; Ullman & Walenski, 2005). Empirically, the findings expand the range of information that speakers are sensitive to, and call for processing models that can capture larger distributional information: While word and bigram frequency effects can be accommodated via links between single words (or a non-symbolic representation of them), frequency effects beyond the bigram—such as the ones mentioned above—necessitate the representation of larger chains of relations, not only between single words but also between word strings of varying sizes.[1] The nature of multi-word frequency effects and the extent to which they affect various aspects of language use is thus of theoretical and empirical significance.

In this paper, we focus on the effect of multi-word frequency on phonetic duration. Several studies have shown that multi-word frequency affects recognition (Arnon & Snider, 2010; Reali & Christiansen, 2007; Siyanova et al., 2011; Tremblay & Baayen, 2010; Tremblay et al., 2011). The findings for production, in particular when looking at phonetic reduction, are more mixed. Speech onset and omission are both affected by multi-word frequency. Speakers start to speak sooner when producing higher frequency sequences (Ellis et al., 2008; Jannsen & Barber, 2012; Tremblay & Baayen, 2010; Tremblay et al., 2011); are more likely to omit linguistic material in more frequent (and predictable) sequences (Jaeger, 2010; Levy & Jaeger, 2007); and use more contractions (*I'm*) in such environments (Frank & Jaeger, 2008). Fewer studies have set out to examine the effect of multi-word frequency (as opposed to bigram frequency, e.g., Jurafsky, Bell, Gregory, & Raymond, 2001) on phonetic duration, and the emerging picture is less clear. While some studies have found that duration is reduced in higher frequency sequences (Aylett & Turk, 2004; Bannard &

Matthews, 2008; Bybee & Schiebmann, 1999), others have failed to find such an effect (Ellis et al., 2008; Tremblay & Tucker, 2011). This pattern is surprising and calls for further investigation given the documented effects of multi-word frequency on other linguistic measures.

Going beyond previous findings, we want to look at phonetic duration to ask if the sensitivity to multi-word statistics is affected by the higher order properties of the sequence in question. So far, studies of multi-word statistics have asked how the surface frequency of the string affects processing. However, it is possible that the impact of multi-word statistics on processing is affected by the cohesiveness of the sequence in other respects. Multi-word sequences can differ in their degree of semantic, syntactic, and prosodic cohesiveness. A sequence such as *don't have to worry* evokes a complete semantic event, forms a syntactic constituent, and can be produced as one intonational phrase. A sequence such as *have to worry about* has similar phrase frequency (2.1 versus 2.7 per million, using the spoken part of the British National Corpus; Davies, 2004), but is less semantically complete, crosses syntactic boundaries, and does not form a complete intonational unit.

Are speakers equally sensitive to multi-word statistics calculated across and within syntactic constituents? Is the effect of multi-word frequency different for words grouped within a prosodic phrase? Examining the effect of higher order properties on the use of multi-word information raises interesting questions both about the psychological reality of such higher order properties and about the way they should (or should not) be integrated into processing models. Finding that the effect of multi-word statistics is attenuated by syntactic, prosodic semantic factors would raise the need to incorporate these properties into emergentist models. In existing implementations, the sensitivity to multi-word statistics (or any other distributional information) is unconstrained by higher order properties; the effect of bigram or trigram frequency is predicted to be the same regardless of whether those words form a cohesive semantic, syntactic, or prosodic unit. Alternatively, finding that speakers are sensitive to multi-word frequency regardless of whether the sequence crosses constituent boundaries would provide support for models that do not expect higher level properties to affect processing. Whatever the outcome, uncovering the full range of multi-word effects and their possible limitations is necessary for the development of adequate processing models.

1.1 Previous findings on the effect of multi-word frequency on phonetic reduction

Many studies have looked at the effect of contextual factors on word duration. Words are phonetically reduced in more predictable semantic and syntactic environments (Gahl & Garnsey, 2004; Jurafsky et al. 2001; Tily et al., 2009), as well as when they appear inside more frequent and predictable sequences (Aylett & Turk, 2004; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Bell et al., 2003). In a seminal study, Bybee and Schiebman (1999) found that *don't* was phonetically reduced in frequently recurring phrases such as *I don't know.* Aylett and Turk (2004) asked how syllable duration is affected by the surrounding context: they show that higher trigram syllable frequency leads to shorter duration in a set of items that included (but was not limited to) three monosyllabic word sequences. Taken together with findings showing effects of multi-word information on omission (Frank & Jaeger, 2008) and voice onset (e.g., Jannsen & Barber, 2012), one would expect to see effects of multi-word frequency on phonetic duration.

However, the few studies that have explicitly asked whether duration is affected by multi-word frequency (as opposed to bigram frequency) have yielded mixed results. Bannard and Matthews (2008) were the first to examine the effect of multi-word frequency (four-word frequency) on phonetic duration. Using a repetition task, they asked two- and three-year-olds to repeat higher and lower frequency phrases that were matched for part frequency and differed only on sequence-frequency (e.g., *a drink of milk*—high versus *a drink of tea*—low). The results showed an effect of multi-word frequency: children at both ages showed shorter durations of the same trigram (*a drink of*) inside higher frequency sequences. Two more recent studies, using a similar elicitation task with adult speakers, failed to find an effect of multi-word frequency on phonetic duration. Ellis et al. (2008) asked native and non-native speakers to produce multi-word sequences of varying frequencies but did not find an effect of *n*-gram on phonetic duration. Their analyses, however, did not control for part frequencies (e.g., the frequency of the unigrams, bigrams, and trigrams), making it harder to isolate the effect of multi-word statistics. Using a more controlled design, Tremblay and Tucker (2011) explored the effect of a number of distributional measures (frequency, probability, and mutual information) on the production of four-word sequences and failed to find an effect of phrase frequency on duration.

What underlies this mixed pattern of results? One possibility is that the effect of multi-word frequency on duration is limited to young children, and is driven by their treatment of the sequence as a more holistic chunk (e.g., Abbot-Smith & Tomasello, 2006). In this case, we would not expect to find such effects in adult speakers. Another, more probable explanation, is that the use of items that were less controlled for part frequency masked the effect in the two adult studies. If so, the effect should be found in adult speakers when using the kind of paired-items used by Bannard and Matthews (2008).

1.2 The current studies

In the current paper, we set out to systematically investigate the effect of multi-word statistics (four-word and three-word frequency) on phonetic duration in adults using a combination of experimental and corpus-based methods. Our first aim is to see whether the effect found by Bannard and Matthews (2008) is found in adults when using items that are (a) well-controlled for part frequency, and (b) have been used previously to document multi-word frequency effects in recognition (Arnon & Snider, 2010). Our second goal is to see if the effect holds also in spontaneous speech. Unlike elicited production, where a sequence is shown in its entirety and then produced, utterances in natural language are longer, are not given to the speaker in advance, and are not fully planned out when speaking begins. Because the studies explicitly looking at multi-word frequency effects on duration all use elicitation tasks, we cannot tell if the effect is dependent on the presentation, and consequent planning, of the sequence as one unit. If so, we would not expect it to appear in spontaneous speech. If, on the other hand, the effect reflects speakers' sensitivity to multi-word information more generally, we would expect the finding to be replicated in spontaneous speech. Our third goal is to go beyond previous literature to explore the possible interaction of higher order properties and multi-word statistics by comparing sequences that form a syntactic constituent with ones that cross syntactic boundaries. We focus on syntactic constituency (as opposed to semantic or prosodic completeness) because it is easier to assess than semantic or prosodic completeness.

We investigate these questions by conducting three studies. The first is an elicited production study in which we compare the duration of identical three-word sequences appearing within a higher or lower frequency phrase of the same syntactic type (e.g., *<u>don't have to</u> worry* versus *<u>don't have to</u> wait*). In the second study we use corpus data to look at the phonetic duration of two three-word structures that differ in their constituency status to test (a) if the effect of multi-word frequency holds in spontaneous speech, and (b) if it is modulated by constituency. The third study is an additional corpus study in which we examine a more diverse range of constructions (constituents and non-constituents) to ensure that the effects are not limited to the two specific constructions examined in the second study.

2 Study 1: Phrase-frequency effects on duration in elicited production

We investigate the effect of phrase frequency on phonetic duration using a phrase-production task where speakers produce a phrase they read on the screen. We compare the duration of the identical trigram in pairs of phrases (all constituents) that differ on phrase frequency (e.g., *don't have to worry* versus *don't have to wait*), but are controlled for part frequency. The use of such matched pairs, and the fact we are looking at the same phonetic material, reduces the variability between the high and low variants and allows us to reduce the collinearity that is often found between the frequencies of different subparts (Tremblay & Tucker, 2011). If adult production, like comprehension, is affected by the frequency of the entire phrase, speakers should show shorter articulation of the same phonetic material inside higher frequency phrases.

2.1 Method

2.1.1 Participants. Thirty-four native-speakers of British English, all students (mean age 20) from the University of Manchester, participated in the experiment. All were native speakers of English and received course credit for participating in the study.

2.1.2 Materials. We selected a subset of the items used by Arnon and Snider (2010) to look at phrase-frequency effects on production. In the original study, each item consisted of two four-word phrases (counting orthographic words) that differed only in the final word (*don't have to worry* versus *don't have to wait*). In each pair, the phrases differed in phrase frequency (high versus 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 to ensure that any difference in processing between the two variants 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, and were rated as describing equally plausible real-world events (norms taken from Arnon & Snider, 2010). The items were constructed using the Switchboard (Godfrey, Holliman, & McDaniel, 1992) and Fisher (Cieri, Miller, & Walker, 2004) corpora that were combined to yield a 20-million-word corpus. Both corpora consist of American English collected from telephone conversations. For the current study, we selected a subset of the items used by Arnon and Snider (2010) that met an additional criteria: Because our participants were speakers of British English we only used items where the frequency criteria for the item-pair held when calculated using the spoken part of the British National Corpora (10 million words; Davies, 2004).

We ended up with 28 constituent item-pairs, each consisting of a higher and lower frequency variant (e.g., *a lot of work* versus *a lot of years*). The mean frequency of the high-frequency phrases (12.04 per million) was higher than that of the low-frequency phrases (2.25 per million), $t(26) = 19.75$, $p < 0.001$. The frequencies of the substrings were no longer perfectly controlled: the frequency of the final word within each pair was higher for low-frequency phrases than for high-frequency ones (high: 1146, low: 1811, $t(26) = -4.75$, $p < 0.001$), while the second trigram was more frequent in the higher frequency phrases compared to the lower frequency ones (high: 24.09, low: 16.03, $t(26) = 3.75$, $p < 0.001$). The frequency of the final bigram did not differ (high: 229, low: 216, $t(26) = .57$, $p = 0.56$). Consequently, all part frequencies were controlled for also in the analyses. The phrases did not differ in the number of letters (high: 13.04, low: 12.82, $t(26) = 1.54$, $p > 0.1$).

2.1.3 Procedure. The experiment was run using DMDX (software developed at Monash University and at the University of Arizona by KI Forster and JC Forster). Participants sat in a quiet room in front of a computer, and completed a phrase-elicitation task. Participants saw four-word phrases on the screen for a fixed amount of time (1700 ms). They were told to produce the phrase as soon as it disappeared from the screen. Participants saw several practice items at the beginning of the experiment. We used a between-subject design to compare durations within an item-pair: each participant saw one variant of each item-pair (38 items in total) to avoid repetition effects (Bard et al., 2000) resulting from producing the same trigram twice (e.g., *don't have to* *worry* and *don't have to* *wait*). Participants were randomly assigned to one of the two experimental lists. The task was divided into two blocks with a short break in between. Each block took about 5 minutes to complete.

2.2 Results

A research assistant blind to the hypotheses of the study measured the duration of the first three words of each response using Praat. Responses whose duration was under 250 ms or over 2 standard deviations from the mean duration were excluded from the analysis. This resulted in the loss of 3.5% of the data. Accuracy of phrase repetition was at ceiling for both conditions (high and low phrase frequency) and reached 99% (incorrect responses were also excluded). We analyzed the results using mixed-effects linear regression models with item and subject as random effects. We used the logged duration of the first three words of each item as the response variable, and ask how duration is affected by phrase frequency (high versus low). We did not use a continuous measure of phrase frequency as a predictor because it is not meaningful in the context of our items: Because the classification into high and low was not consistent across items (e.g., high is over 10 per million) but only within a pair (a high-frequency item is only high relative to its low-frequency pair), the use of continuous frequency was not informative.

In addition to our variable of interest, we used several control variables. We added the logged frequencies of all the substrings that differed between the high and low variants (second trigram, third bigram, and fourth unigram). We also controlled for the average speech rate of the speaker (calculated over all their responses in the experiment), the expected duration of the first trigram (calculated over all productions of that trigram in the experiment), and the length in syllables of the phrase. These controls enabled us to examine the additional variance explained by phrase frequency. There was no need to control for the length (in syllables) of the first trigram, since there was very little variance between items (all were either three or four syllables long) but we did control for list (which of the two experimental lists the participant saw). We included subject and item as random effects.

In all analyses, we checked for collinearity between the fixed effects (e.g., between the frequency of the final trigram and the frequency of the final bigram) and reduced it when necessary 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 0.22.

As predicted, phonetic duration was affected by phrase frequency: durations were shorter inside higher frequency phrases. The model showed a significant effect of phrase frequency: duration was shorter in higher frequency variants when controlling for substring frequency ($\beta = -0.017$, $SE = 0.001$, pMCMC < 0.05). We used model comparisons to see if our variable of interest (phrase frequency) significantly improves the model by comparing a model with the variable of interest to one without the variable (but with all other fixed and random effects). We do this comparison using the

Table 1. Elicited production: Regression coefficients and *p* values for constituent items (variable of interest in bold).

Variable	Coef. β	SE(β)	*p*MCMC	*p*
Intercept	–0.3533	(0.1904)	0.0342	0.0373
Freq-type-lo	0.01694	(0.0082)	0.0418	0.0406
Fourth unigram	0.00673	(0.00512)	0.1984	0.2105
Third bigram	0.0003	(0.0019)	0.8372	0.9577
Second trigram	–0.0011	(0.0026)	0.5676	0.6524
Rate of speech (log)	0.1791	(0.0765)	0.0018	0.0195
Expected duration (log)	0.9631	(0.0688)	0.0001	0.0000
Number of syl. (log)	0.0502	(0.0401)	0.1918	0.2105

non-residualized predictors (see Jaeger, 2010, for more details). Phrase frequency was a significant predictor using model comparisons ($\chi^2 = 5.54$, $p < 0.05$). In addition to phrase frequency, durations were affected by individual speech rate and expected duration. Not surprisingly, slower individual speech rates led to longer durations ($\beta = 0.17$, SE = .18, *p*MCMC < 0.01, $p < 0.05$ in model comparisons), as did longer expected durations ($\beta = 0.96$, SE = .07, *p*MCMC < 0.001, $p < 10^{-4}$ in model comparisons). No other factors were significant (all *p* values > 0.18, see Table 1).

2.3 Discussion

The results of the first study document multi-word frequency effects on phonetic duration in adults. These findings replicate the ones found for children (Bannard & Matthews, 2008) and suggest that the lack of an effect in previous elicitation studies with adults (Ellis et al., 2008; Tremblay & Tucker, 2011) was driven by insufficient control of part frequencies or study-specific properties (e.g., the inclusion of many different frequency measures in one analysis). The effect of multi-word frequency was not limited to highly collocated sequences (e.g., *I don't know*) and was found for sequences of varying frequencies: these findings extend our understanding of the range of information adult speakers are sensitive to in production.

Having established that phonetic duration in adults *is* affected by multi-word frequency, we can now ask two additional questions. The first is whether the effect holds in spontaneous speech when the sequence in question is not presented and planned as one unit, as is suggested given findings on syllable duration (Aylett & Turk, 2004) and contractions (Frank & Jaeger, 2008). The second—and novel question—is if multi-word frequency affects phonetic duration in the same way for constituents and non-constituents. This question is important in exploring the limitations on the use of multi-word information in processing and how it may interact with higher order properties such as constituency.

3 Study 2: Phrase-frequency effects on duration in spontaneous speech: comparing two structures

In this study, we look at the effect of multi-word frequencies and constituency on phonetic duration in spontaneous speech using the Switchboard corpus. This is, to our knowledge, the first test of the effect of multi-word frequencies on duration in spontaneous speech. The use of spontaneous speech has several additional advantages. Spontaneous speech varies across several axes—the word frequencies of each word in each sequence are not kept constant, the words used have different phonological properties, and the multi-word sequence may occur in the beginning, middle, or end of prosodic units. Finding multi-word frequency effects across such varied stimuli would increase the validity and generality of the findings. More importantly, the use of spontaneous speech allows us to ensure that all sequences were produced with natural prosody and to circumvent the limitation of using an elicitation task to generate non-constituents—producing a four-word sequence as a standalone phrase (as in the elicitation task) is more unnatural for non-constituents than for constituents (compare, for instance, *as far as I* to *don't have to worry*).

We contrast two three-word syntactic structures in which the second word is a function word. The first structure consists of subject–auxiliary–verb sequences (e.g., *everybody was trying*)—which are not seen as one complete constituent (Marantz, 1981). The second structure consists of verb–determiner–noun sequences (e.g., *saw the boy*), which are taken to be one constituent. If constituency plays a significant part in the retrieval and production of multi-word sequences, we should see a smaller (or no) effect of frequency on the duration of trigrams in non-constituent sequence compared to trigrams in constituent sequences.

3.1 Materials and method

We used the parsed section of the Switchboard corpus and aligned each word with its duration and syntactic properties. Each word was matched with (a) its part-of-speech tag from the parsed section of the corpus, (b) its duration from the time-aligned section of the corpus (see Deshmukh, Ganapathiraju, Gleeson, Hamaker, & Picone, 1998, for full details on how the time measures were extracted), and (c) its number of syllables using the CMU dictionary (Weide, 1998). We calculated rate of speech as the number of lexical segments per second (including omitted and reduced segments). To control for phrase-final lengthening, we had to identify which of the sequences appeared at an end of a phrase. Each sequence was matched with its location in a prosodic phrase and coded as either phrase-final or not phrase-final: phrases were considered to be continuous speech, allowing gaps of less than 0.1 s.

We used the parsed section of the corpus to extract items from the two syntactic structures. For the constituent items, we extracted all verb–determiner–noun sequences in which verbs were followed by a direct object consisting of a determiner and a single-word noun, such as *find a way, doing the work*. We excluded sequences in which the direct objects were *a lot, a couple*, and *a few*, since these noun phrases function as adverbials in such constructions, rather than as nouns. We also excluded sequences where the duration of one of the words was missing. We ended up with 2366 observations.

For the non-constituent items (noun–verb–verb), we extracted sequences in which a single-word noun was followed by an auxiliary and a verb. We counted pronouns as nouns, and excluded determiners when they appeared before the noun. The auxiliaries were restricted to *do, have, be* and modal verbs such as *will* (the full list is *have, are, was, is, were, has, had, am, been, being, got, get, be, did, does, do, would, should, will, can, could, must, might, may*). We only used sequences where the verb was followed by at least another word within the verb phrase—we did this to ensure that the three-word sequence did not appear as a standalone constituent. Examples included *people will take, someone is taking, they have taken*. We excluded sequences where the duration of one of the words was missing. We ended up with 2290 observations.

Unlike the previous study, the corpus study does not allow us to compare multiple instances of carefully selected pairs. Instead, we controlled for the log frequencies of every word and two-word and three-word sequences. We used word and sequence frequencies from the corpus without smoothing, and without removing low-frequency items. Word and sequence frequencies are highly collinear. Therefore, sequence log frequencies were residualized based on their respective substrings: the frequency of the sequence of the first two words (first bigram) was residualized using the frequencies of the first and second word, the frequency of the second and third word sequence (second bigram) was residualized using the frequencies of the second and third word and the frequency of three-word sequences (trigram) was residualized using the frequencies of every word in isolation, as well as the frequencies of the sequences of the first two words and the last two words. We used the Fisher and Switchboard corpora to calculate word frequencies and sequence frequencies.

To reduce variability further, we used the Switchboard corpus to calculate the expected duration for each word—using the mean of the logged duration of every occurrence of that word when it appeared with the same part-of-speech tag as the one it had in our data. We calculated the expected duration of the three-word sequences by summing the expected duration of the three words, and used this measure as an additional control in our analyses. Finally, we calculated the duration of the three-word sequences by summing the observed duration of the three words.

3.2 Results

We excluded sequences whose duration was more or less than 2.5 standard deviations from the mean duration of all sequences. This resulted in loss of 2% of the data, leaving us with 2515 observations for constituents and 2289 for non-constituents. For both data sets, we used a mixed-effects linear regression to estimate the log duration of the entire three-word sequence based on the expected duration of the three words. We used several phonological and phonetic controls: the speaker's logged rate of speech, a binary variable that indicated whether the word sequence was at the end of a prosodic phrase (to control for phrase-final lengthening: this was only relevant for the constituent data, since the non-constituents were extracted as to not be at the end of a phrase), and the log number of syllables in the sequence. To control for the frequencies of the subparts, we included the log frequencies of the three words, and the residual frequencies of the sequence of the first two words and the last two words. The identity of the speaker was used as a random effect. We report the results for the constituents and non-constituents separately before reporting the conjoined model.

3.2.1 Constituents. Paralleling the results of the experiment, higher multi-word frequency led to shorter durations of the three-word sequence in spontaneous speech ($\beta = -0.0048$, SE = 0.0017, pMCMC < 0.01, $p < 0.01$ in model comparison). The significance of sequence-frequency was not dependent on the inclusion of speech rate, expected duration, end of phrase, or number of syllables (although their inclusion significantly improved the model, all p values < 0.001 in model comparisons). Not surprisingly, high speech rate led to shorter durations ($\beta = -0.6365$, SE = 0.0157, pMCMC $< 10^{-3}$, model comparison $p < 10^{-4}$), while having a higher number of syllables ($\beta = 0.1157$, SE = 0.0162, pMCMC < 0.001, model comparison $p < 10^{-4}$), being at the end of a phrase ($\beta = 0.0328$, SE = 0.0061, pMCMC < 0.001, model comparison $p < 10^{-4}$), and having a longer expected duration ($\beta = 0.8126$, SE = 0.0177, pMCMC < 0.001 model comparison $p < 10^{-4}$) all led to longer sequence duration.

Duration was also significantly affected by part frequencies. It was shorter when the frequencies of the third word and first bigram were higher (third word: $\beta = -0.0026$, SE = 0.0011, pMCMC < 0.05, $p > 0.35$ in model comparison;[2] first bigram: $\beta = -0.0041$, SE = 0.0025, pMCMC < 0.05, model comparison $p > 0.5$). The frequency of the second unigram had an inverse effect on duration: the duration of the sequence was longer when the second word was more frequent (second word:

Table 2. First corpus study: Regression coefficients and *p* values for constituent items (variable of interest in bold).

Variable	Coef. β	SE(β)	z	*p*MCMC	*p*
Intercept	1.5311	(0.0445)	34.4390	0.0001	0.0000
Sequence-frequency	–0.0048	(0.0017)	–2.8370	0.0092	0.0046
First word unigram	0.0019	(0.0011)	1.7438	0.0610	0.0813
Second word unigram	0.0060	(0.0020)	3.0028	0.0020	0.0027
Third word unigram	–0.0026	(0.0011)	–2.4161	0.0146	0.0158
First bigram	–0.0041	(0.0025)	–1.6855	0.0456	0.0920
Second bigram	–0.0015	(0.0021)	–0.7393	0.5602	0.4598
Rate of speech (log)	–0.6365	(0.0157)	–40.6234	0.0001	0.0000
Expected duration (log)	0.8126	(0.0177)	45.8933	0.0001	0.0000
Number of syl. (log)	0.1157	(0.0162)	7.1549	0.0001	0.0000
End of phrase	0.0328	(0.0061)	5.4213	0.0001	0.0000

$\beta = 0.006$, SE = 0.002, *p*MCMC < 0.01, model comparison $p < 0.01$). This (unexpected) effect probably reflects an idiosyncrasy of the data set: because all of the sequences had a verb–determiner–noun structure, there were few word types (only 16) that appeared as the second word. The lengthening effect seems to be driven by the two most frequent determiners (*a* and *the*), which appear with a more diverse set of nouns that the other determiners. None of the other effects were significant (see Table 2).

3.2.2 Non-constituents. The results for the non-constituents paralleled the ones found for constituents: Higher multi-word frequency led to shorter durations ($\beta = -0.0076$, SE = 0.0032, *p*MCMC < 0.05, $p < 0.05$ in model comparison). Higher rate of speech led to shorter durations ($\beta = -0.7942$, SE = 0.0224, *p*MCMC < 0.001, model comparison $p < 10^{-4}$), while higher number of syllables ($\beta = 0.1338$, SE = 0.0367, *p*MCMC < 0.01, model comparison $p > 0.001$) and longer expected duration ($\beta = 0.8424$, SE = 0.0268, *p*MCMC < 0.001, model comparison $p > 10^{-4}$) led to longer actual durations. Here, also, duration was affected by part frequency. Duration was shorter when the frequencies of the third word and second bigram were higher (third word: $\beta = -0.0064$, SE = 0.0021, *p*MCMC < 0.01, model comparison $p > 0.6$; second bigram: $\beta = -0.0098$, SE = 0.0028, *p*MCMC < 0.001, model comparison $p > 0.25$). None of the other effects were significant (see Table 3).

3.2.3 A combined model. The effect of multi-word frequency was significant for both constituents and non-constituents: in both cases, higher multi-word frequency led to shorter durations. To examine whether the size of the effect differs for constituents and non-constituents, we ran a third combined model, where the main objective was to test if there was a significant interaction for type and multi-word frequency. We used the same controls used in the two separate models (part frequency, log(speech rate), log(number of syllables), expected duration, and end-of-phrase), and added all the two-way interactions between type (constituent versus non-constituent) and the other control variables to see if the two data sets differed on additional dimensions except the one we were testing (sequence-frequency X type). We discuss only the interactions and effects that were significant in the model and in model comparisons (for the full model, see Table 4).

The effect of sequence-frequency remained significant ($\beta = -0.0049$, SE = 0.002, *p*MCMC < 0.05, model comparison $p < 0.01$), and there was a significant effect of type: constituents were

Table 3. First corpus study: Regression coefficients for non-constituent items (variables of interest in bold).

Variable	Coef. β	SE(β)	z	pMCMC	p
Intercept	1.8970	(0.0727)	26.0772	0.0001	0.0000
Sequence-frequency	**–0.0076**	**(0.0032)**	**–2.4048**	**0.0190**	**0.0163**
First word unigram	–0.0004	(0.0023)	–0.1768	0.8126	0.8597
Second word unigram	0.0011	(0.0033)	0.3416	0.7958	0.7327
Third word unigram	–0.0064	(0.0021)	–3.0368	0.0034	0.0024
First bigram	–0.0009	(0.0037)	–0.2360	0.9010	0.8134
Second bigram	–0.0098	(0.0028)	–3.4381	0.0014	0.0006
Rate of speech (log)	–0.7942	(0.0224)	–35.4114	0.0001	0.0000
Expected duration (log)	0.8424	(0.0268)	31.4849	0.0001	0.0000
Number of syl. (log)	0.1338	(0.0367)	3.6436	0.0004	0.0003

Table 4. First corpus study: Regression coefficients for combined model (excluding non-significant interaction terms, variables of interest in bold).

Variable	Coef. β	SE(β)	z	pMCMC	p
Intercept	1.5344	(0.0528)	29.0709	0.0001	0.0000
Sequence-frequency	**–0.0049**	**(0.0020)**	**–2.4264**	**0.0260**	**0.0153**
Non-constituent	0.3513	(0.0825)	4.2589	0.0001	0.0000
First word unigram	0.0018	(0.0013)	1.3673	0.1118	0.1716
Second word unigram	0.0059	(0.0024)	2.4531	0.0098	0.0142
Third word unigram	–0.0027	(0.0013)	–2.0967	0.0384	0.0361
First bigram	–0.0041	(0.0029)	–1.4113	0.0954	0.1582
Second bigram	–0.0014	(0.0025)	–0.5824	0.6158	0.5603
Rate of speech (log)	–0.6376	(0.0186)	–34.2886	0.0001	0.0000
Expected duration (log)	0.8130	(0.0210)	38.6702	0.0001	0.0000
End of phrase	0.0326	(0.0072)	4.5142	0.0001	0.0000
Number of syl. (log)	0.1154	(0.0192)	5.9976	0.0001	0.0000
Sequence-frequency: Non-constituent	–0.0025	(0.0034)	–0.7248	0.3904	0.4686
Non-constituent:First word unigram	–0.0021	(0.0024)	–0.8755	0.2952	0.3814
Non-constituent:Second word unigram	–0.0044	(0.0038)	–1.1583	0.1540	0.2468
Non-constituent:Third word unigram	–0.0036	(0.0022)	–1.5932	0.0948	0.1112
Non-constituent:First bigram	0.0029	(0.0043)	0.6779	0.3164	0.4979
Non-constituent:Second bigram	–0.0085	(0.0035)	–2.4216	0.0158	0.0155
Non-constituent:Rate of speech (log)	–0.1511	(0.0269)	–5.6108	0.0001	0.0000
Non-constituent:Expected duration (log)	0.0318	(0.0315)	1.0109	0.3432	0.3121
Non-constituent:Number of syl. (log)	0.0156	(0.0375)	0.4161	0.6578	0.6773

longer than non-constituents overall ($\beta = 0.3513$, SE = 0.0825, *p*MCMC < 0.001, model comparison $p < 10^{-4}$). The interaction of sequence-frequency and constituency, however, was not significant: we saw no difference in the effect of multi-word frequency on the duration of constituents and non-constituents ($p > 0.4$ in model comparisons). Two other interactions were significant: (1) type interacted with speech rate: higher speech rate reduced the duration of the sequence more in non-constituents than in constituents ($\beta = –0.1511$, SE = 0.0269, *p*MCMC < 0.001, model comparison

$p < 10^{-4}$); and (2) type interacted with the frequency of the second bigram: higher second bigram frequency reduced the duration of the sequence more for non-constituents ($\beta = -0.0085$, SE = 0.0035, *p*MCMC < 0.05 $p > 0.1$).

As in the separate models, higher rate of speech led to shorter durations ($\beta = -0.6376$, SE = 0.0186, *p*MCMC < 0.001, model comparison $p < 10^{-4}$), while higher number of syllables (β = 0.1154, SE = 0.0192, *p*MCMC < 0.001, model comparison $p < 10^{-4}$), being at the end of a phrase (β = 0.0326, SE = 0.0072, *p*MCMC < 0.001, model comparison $p < 10^{-4}$), and having longer expected duration β = 0.813, SE = 0.021, *p*MCMC < 0.001, model comparison $p < 10^{-4}$) all led to longer actual durations. Two of the control variables were significant: High second word frequency increased the duration of the sequence (β = 0.0059, SE = 0.0024, *p*MCMC < 0.05, $p < 0.05$ in model comparison), and high third word frequency decreased the duration of the sequence (β = –0.0027, SE = 0.0013, *p*MCMC < 0.05, $p > 0.58$ in model comparison). To summarize, type did not moderate the effect of multi-word frequency.

3.3 Discussion

The corpus findings mirrored the results found in the experimental task. Phonetic duration in spontaneous speech was affected by multi-word frequency. Over a diverse set of items, durations were shorter for higher frequency sequences even when all substring frequencies were controlled for. These findings join previous work in documenting the effect of contextual factors on phonetic reduction (e.g., Aylett & Turk, 2004; Bell et al., 2009; Jurafsky et al., 2001). They further show that the effect holds in spontaneous speech and cannot be attributed to the ‘chunked’ presentation of stimuli in elicitation tasks. Phonetic durations are affected not only by features of the word itself, or the words appearing directly after and before it, but also by the properties of the larger linguistic context: at a minimum, the findings reinforce the effect of trigram frequency on production (Aylett & Turk, 2004; Bell et al., 2009; Frank & Jaeger, 2008).

Importantly, the results reveal parallels between constituents and non-constituents in the effect of multi-word frequency: duration was shorter in higher frequency sequences regardless of constituency status and there was no evidence that the effect differed in magnitude between the two types. To further examine this novel finding, and to ensure it is not limited to the two specific constructions we examined in Study 2 (*noun–determiner–verb* and *noun–auxiliary–verb*), we conducted a third corpus study in which we examined duration in a variety of three-word constituent and non-constituent structures all extracted from the same post-verbal position. If the effect of multi-word frequency is indeed similar in constituents and non-constituents, we should find it (again) for both types across multiple constructions. Looking at a wider range of constructions will also serve to extend the generality and validity of the effect we found and ensure it is not limited to any particular construction.

4 Study 3: Phrase-frequency effects on duration in spontaneous speech: comparing multiple structures

In this study, we contrast three-word sequences extracted from the same post-verbal position: constituents were sisters of the verb in a verb phrase, such as *once a year* in *pruning once a year, from New Mexico* in *she was from New Mexico*, or *what I mean* in *that's what I mean*, while non-constituents were sequences taken from across two sisters in a verb phrase, such as *this to you* in *doing this to you, him that way* in *perceived him that way*, or *up using it* in *ended up using it*. The constituents and non-constituents were similar in their syntactic position in the sentence and in the

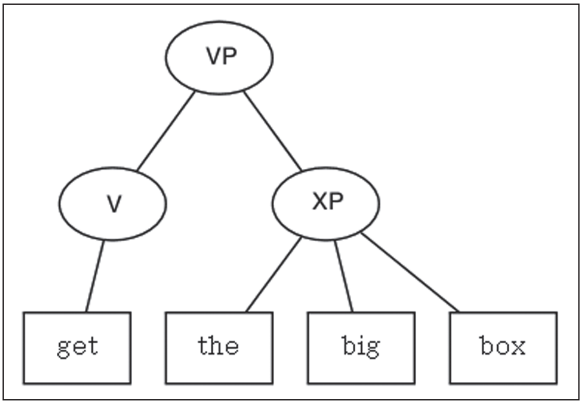

Figure 1. Tree structure for constituents (Study 3).

diversity of the words and part-of-speech appearing in them. Because, unlike the previous study, we did not limit ourselves to a particular sequence of word types (e.g., *verb–determiner–noun*), the resulting set of items included a wide range of words, phrase types (verb phrases, noun phrases, prepositional phrases), and constructions. If constituency plays a significant part in the retrieval and production of multi-word sequences, we should see a smaller (or no) effect of frequency on the duration of trigrams in non-constituent sequence as compared with constituent sequences.

4.1 Materials and method

As in the previous study, we used the parsed section of the corpus to extract every four-word verb phrase in the corpus. Following the procedure used in Study 2, each word was matched with its (a) part of speech, (b) duration (using the time-aligned section of the corpus), (c) expected duration based on all appearances in the corpus, (d) number of syllables, and (e) whether it appeared at the end of a prosodic phrase (see previous study for more details) .

Constituents were taken to be any sister of the verb in a verb phrase, provided that the verb had only one sister: *(get) [the big box], (stayed) [for a while]*, as illustrated in Figure 1. Non-constituents were sequences straddling two sisters of the verb in the verb phrase: *(gave) [the box] [away], (hand) [it] [to him]*, as illustrated in Figure 2. Other verb phrase structures were discarded. We excluded sequences where the duration of one of the words was missing. We ended up with 3746 observations for constituents, and 1619 observations for non-constituents. We used the exact same statistical procedure, controls, and frequency calculations as in Study 2.

4.2 Results

We excluded sequences whose duration was more or less than 2.5 standard deviations from the mean duration of all sequences. This resulted in loss of 1.6% of the data, leaving us with 3684 observations for constituents and 1596 for non-constituents. For both data sets, we used a mixed-effects linear regression to estimate the log duration of the entire three-word sequence based on the expected duration of those three words. We used the same frequency and speech-related controls as in Study 2.

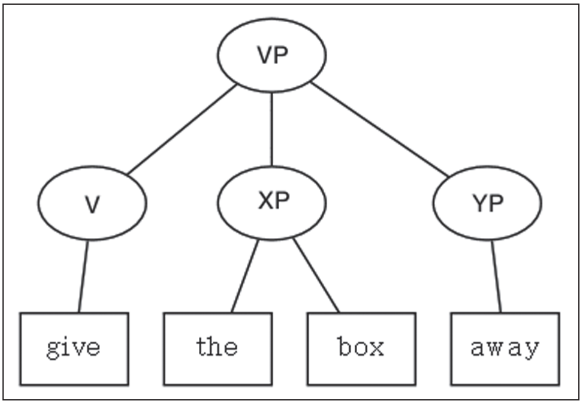

Figure 2. Tree structure for non-constituents (Study 3).

Table 5. Second corpus study: Regression coefficients for constituents (variables of interest in bold).

Variable	Coef. β	SE(β)	*z*	*p*MCMC	*p*
Intercept	1.7242	(0.0416)	41.4072	0.0001	0.0000
Sequence-frequency	–0.0038	(0.0016)	–2.3106	0.0214	0.0209
First word unigram	–0.0012	(0.0012)	–1.0850	0.2742	0.2780
Second word unigram	0.0002	(0.0010)	0.2013	0.8598	0.8404
Third word unigram	–0.0007	(0.0009)	–0.7467	0.4550	0.4553
First bigram	–0.0010	(0.0014)	–0.7070	0.4800	0.4796
Second bigram	–0.0040	(0.0012)	–3.4271	0.0002	0.0006
Rate of speech (log)	–0.7043	(0.0148)	–47.4419	0.0001	0.0000
Number of syl. (log)	0.1009	(0.0130)	7.7646	0.0001	0.0000
End of phrase	0.0366	(0.0054)	6.8201	0.0001	0.0000
Expected duration (log)	0.7520	(0.0111)	67.8414	0.0001	0.0000

4.2.1 Constituents. Paralleling the results of the experiment and the previous corpus study, higher multi-word frequency led to shorter durations of the three-word sequence in spontaneous speech ($\beta = -0.0038$, SE = 0.0016, *p*MCMC < 0.05, $p < 0.05$ in model comparison). The significance of sequence-frequency was not dependent on the inclusion of any of the control variables. Not surprisingly, high speech rate led to shorter durations ($\beta = -0.7043$, SE = 0.0148, *p*MCMC < 0.001, model comparison $p < 10^{-4}$), while being at the end of a phrase ($\beta = 0.0366$, SE = 0.0054, *p*MCMC < 0.001, model comparison $p < 10^{-4}$), having a higher number of syllables ($\beta = 0.1009$, SE = 0.013, *p*MCMC < 0.001, model comparison $p < 10^{-4}$), and having higher expected durations ($\beta = 0.752$, SE = 0.0111, *p*MCMC < 0.001, model comparison $p < 10^{-4}$) all led to longer durations. Duration was also significantly affected by part frequencies. It was shorter when the frequency of the second bigram was higher ($\beta = -0.004$, SE = 0.0012, *p*MCMC < 0.001, $p > 0.3$ in model comparison). All other variables were not significant (see Table 5).

4.2.2 Non-constituents. As was the case in Study 2, and in line with the findings for constituents, there was an effect of multi-word frequency on phonetic duration in non-constituents: higher frequency sequences were shorter ($\beta = -0.0051$, SE = 0.0026, *p*MCMC < 0.05, marginally significant

Table 6. Second corpus study: Regression coefficients for non-constituents (variables of interest in bold).

Variable	Coef. β	SE(β)	z	pMCMC	p
Intercept	1.8546	(0.0609)	30.4568	0.0001	0.0000
Sequence-frequency	–0.0051	(0.0026)	–1.9417	0.0508	0.0523
First word unigram	–0.0018	(0.0017)	–1.0655	0.2638	0.2868
Second word unigram	0.0017	(0.0016)	1.0803	0.2902	0.2802
Third word unigram	–0.0029	(0.0016)	–1.8776	0.0686	0.0606
First bigram	–0.0034	(0.0021)	–1.5905	0.1124	0.1119
Second bigram	–0.0067	(0.0019)	–3.5787	0.0001	0.0004
Rate of speech (log)	–0.7586	(0.0216)	–35.1347	0.0001	0.0000
Number of syl. (log)	0.1123	(0.0190)	5.9166	0.0001	0.0000
End of phrase	0.0211	(0.0079)	2.6672	0.0078	0.0077
Expected duration (log)	0.7489	(0.0170)	44.1181	0.0001	0.0000

Table 7. Second corpus study: Regression coefficients for the combined model (excluding non-significant interactions, variables of interest in bold).

Variable	Coef. β	SE(β)	z	pMCMC	p
Intercept	1.7086	(0.0414)	41.2975	0.0001	0.0000
Sequence-frequency	–0.0038	(0.0016)	–2.3718	0.0188	0.0177
Non-constituent	0.1263	(0.0737)	1.7138	0.1224	0.0866
First word unigram	–0.0015	(0.0011)	–1.3442	0.2658	0.1789
Second word unigram	0.0000	(0.0009)	0.0057	0.8636	0.9955
Third word unigram	–0.0007	(0.0009)	–0.7902	0.4508	0.4294
First bigram	–0.0011	(0.0014)	–0.7693	0.4752	0.4418
Second bigram	–0.0039	(0.0011)	–3.4298	0.0010	0.0006
Rate of speech (log)	–0.6964	(0.0148)	–47.0818	0.0001	0.0000
Number of syl. (log)	0.0988	(0.0129)	7.6863	0.0001	0.0000
End of phrase	0.0317	(0.0044)	7.1457	0.0001	0.0000
Expected duration (log)	0.7530	(0.0110)	68.6933	0.0001	0.0000
Sequence-frequency:Non-constituent	–0.0013	(0.0031)	–0.4034	0.6546	0.6867
Non-constituent:First word unigram	–0.0001	(0.0021)	–0.0516	0.7774	0.9589
Non-constituent:Second word unigram	0.0017	(0.0019)	0.8979	0.4284	0.3693
Non-constituent:Third word unigram	–0.0022	(0.0018)	–1.1928	0.2592	0.2330
Non-constituent:First bigram	–0.0025	(0.0026)	–0.9781	0.3374	0.3281
Non-constituent:Second bigram	–0.0027	(0.0022)	–1.2002	0.2162	0.2301
Non-constituent:Rate of speech (log)	–0.0563	(0.0264)	–2.1350	0.0538	0.0328
Non-constituent:Number of syl. (log)	0.0119	(0.0233)	0.5119	0.6758	0.6088
Non-constituent:Expected duration (log)	–0.0026	(0.0205)	–0.1280	0.9280	0.8982

in model comparison $p = 0.51$). The effect was not dependent on the inclusion of any of the control variables. Like in the constituent data, and like in the previous study, higher rate of speech led to shorter durations ($\beta = –0.7586$, SE = 0.0216, pMCMC < 0.001, model comparison $p < 10^{-4}$), while longer expected durations ($\beta = 0.7489$, SE = 0.017, pMCMC < 0.001, model comparison $p < 10^{-4}$), having a higher number of syllables ($\beta = 0.1123$, SE = 0.019, pMCMC < 0.001, model comparison

$p < 10^{-4}$), and being at the end of phrase ($\beta = 0.0211$, SE = 0.0079, pMCMC < 0.01, model comparison $p < 0.01$) led to longer actual durations. The frequency of the second bigram was also significant: durations were shorter for higher frequency bigrams ($\beta = -0.0067$, SE = 0.0019, pMCMC < 0.001, $p > 0.14$ in model comparison). All other factors were not significant (see Table 6).

4.2.3 A combined model. As in Study 2, we wanted to see whether the size of the effect differs for constituents and non-constituents. To do so, we ran a third combined model, where the main objective was to test if there was a significant interaction for type and multi-word frequency. We used the exact same controls and procedure as in Study 2. We discuss only the interactions and effects that were significant in the model and in model comparisons (for the full model, see Table 7).

The effect of sequence-frequency remained significant ($\beta = -0.003$, SE = 0.001, pMCMC < 0.05, model comparison $p < 0.01$). The effect of type was not significant ($\beta = 0.126$, SE = 0.073, pMCMC > 0.1, model comparison $p > 0.09$). The interaction of sequence-frequency and constituency was also not significant: we saw no difference in the effect of multi-word frequency on the duration of constituents and non-constituents ($p > 0.6$ in model comparisons). One other interaction were significant: (1) type interacted with speech rate: higher speech rate reduced the duration of the sequence more in non-constituents than in constituents ($\beta = -0.056$, SE = 0.026, pMCMC = 0.055, model comparison $p < 0.05$).

As in the separate models, higher rate of speech led to shorter durations ($\beta = -0.6964$, SE = 0.014, pMCMC < 0.001, model comparison $p < 10^{-4}$), while higher number of syllables ($\beta = 0.098$, SE = 0.012, pMCMC < 0.001, model comparison $p < 10^{-4}$), being at the end of a phrase ($\beta = 0.031$, SE = 0.004, pMCMC < 0.001, model comparison $p < 10^{-4}$), and having longer expected duration ($\beta = 0.753$, SE = 0.011, pMCMC < 0.001, $p < 10^{-4}$) all led to longer actual durations. Duration was also affected by the second bigram frequency: higher second bigrams led to shorter durations ($\beta = -0.003$, SE = 0.001, pMCMC < 0.05, $p > 0.2$ in model comparison). To summarize, type did not moderate the effect of multi-word frequency in Study 3.

4.3 Discussion

In Study 3, we examined the effect of multi-word frequency and constituency on phonetic duration in spontaneous speech by looking at a broader range of structures. We wanted to see if the lack of a difference between constituents and non-constituents in the effect of frequency on duration was 'real' or if it was an artifact of the specific set of sequences we examined. In this study also we found a significant effect of multi-word frequency on duration for both constituents and non-constituents. In both data sets, higher sequence-frequency led to shorter durations and there was no indication the effect was different in magnitude for constituents and non-constituents. Taken together, the findings point to a robust effect of multi-word frequency on duration regardless of constituency, and demonstrate that they are not limited to the particular constructions looked at in the first study.

5 General discussion

Using a combination of experimental and corpus-based research we set out to investigate the effect of multi-word statistics on production—specifically on phonetic duration. We asked (a) whether duration is reduced in higher frequency sequences, (b) whether the effect is found in both elicited and spontaneous speech, and (c) whether constituency modulates this effect. The results of the three studies (one of elicited production and two corpus studies) provide an affirmative question to the first two questions and a negative answer to the third. The phonetic duration of three- and four-word sequences was affected by sequence-frequency: productions were shorter in higher frequency sequences, indicating that production is sensitive to larger distributional patterns. The effect was found across a range of frequencies (not only when looking at highly collocated items) and cannot be attributed to substring frequency—the pairs of items were matched for substring frequencies in the experimental manipulation, and substring frequencies were controlled for in all analyses. The effect was found in both elicited and spontaneous speech when controlling for a host of variables known to affect phonetic duration.

Interestingly, the effect was found for both constituents and non-constituents: phonetic duration was reduced for higher frequency sequences regardless of constituency. How should we interpret these findings? Why is it that even though constituents form a more cohesive unit, and are more likely to be part of the same prosodic phrase, they are similarly affected by surface frequencies? One possibility is that the effect of frequency on duration happens without access to higher level properties: information about constituency may no longer be available (or relevant) during articulation. This interpretation is compatible with models of processing that do not encode constituency as a separate property, but instead expect it to emerge as a function of the distributional characteristics of the linguistic element in question (e.g., McClelland et al., 2010). While interesting, this interpretation would have to be reconciled with findings showing that syntactic probability (higher order information) does affects single-word duration (Gahl & Garnsey, 2004; Tily et al., 2009).

Alternatively, we did not find a difference between constituents and non-constituents because there are other relevant differences between the two types that we did not model. This interpretation receives some support from the intriguing interaction of speech rate and type found in the corpus studies: while high speech rate reduces the duration of all segments, its effect on non-constituents is greater. It is also consistent with the pattern of results found in the literature. While not explicitly targeted, existing studies have looked at both constituents and non-constituents. The general pattern seems to be that studies that looked at the effect of multi-word statistics within a syntactic constituent report finding such effects (Arnon & Snider, 2010; Bannard & Matthews, 2008; Bybee & Schiebmann, 1999; Reali & Christiansen, 2007), while studies that use a mix of constituents and non-constituents result in more mixed findings. Some such studies either fail to find an effect of multi-word frequency (Ellis et al., 2008; Tremblay & Tucker, 2011) or find a significant interaction between phrase type and multi-word frequency (Tremblay & Baayen, 2010), while others find an effect across constituent and non-constituent items (Frank & Jaeger, 2008). Further work is needed to uncover the dimensions that affect duration in constituents and non-constituents and their possible interaction with multi-word frequency.

The results reported here advance our understanding of the way multi-word statistics affect processing in several ways. Firstly, they confirm that multi-word frequency affects phonetic duration in adults, and in doing so expand the range of linguistic behaviors sensitive to larger distributional information. Secondly, they show the effect holds in spontaneous speech across a wide range of constructions. Finally, the results open up a novel set of questions about the significance (or lack thereof) of higher level properties during online production.

5.1 Implications for models of language

Part of the motivation for asking how multi-word statistics affect processing stems from the different predictions made on the topic by traditional and emergentist models of language. Phrase-frequency effects—such as the ones reported here—are hard to accommodate within a words-and-rules approach that maintains a distinction between forms stored in the lexicon and ones generated by grammar (e.g., Pinker, 1999). Because frequency effects are thought to be a property of memorized forms (Pinker & Ullman, 2002), such models would either have to assume that all multi-word sequences are stored, or adopt a dual-access model where all forms can be both stored and generated (Ullman & Walenski, 2005). Both solutions would allow the models to capture phrase-frequency effects but at the cost of undermining the empirical distinction between 'stored' and 'computed' forms: a distinction that lies at the heart of their theoretical position. The fact that both constituents and non-constituents showed frequency effects poses an additional challenge, since non-constituents are not perceived as units in such accounts.

The effect of multi-word statistics on production supports the basic underpinning of emergentist models by demonstrating parallels in the processing of words and larger sequences and undermining an empirical distinction between atomic and complex forms. The lack of a difference between constituents and non-constituents is compatible with existing emergentist models that do not encode higher level properties but instead capture online processing effects using surface-level features (Baayen & Henrix, 2011; Elman, 2009; McClelland et al., 2010).

The findings also have implications for models of production. Even though there is evidence that a certain degree of planning happens at the clause level (Miller & Weinert, 1998; Smith & Wheeldon, 1999), classic models of production (Levelt, Roelofs, & Meyer, 1999) conceptualize speech as involving a move from the message, through the selection of lemmas from the lexicon, to the grammatical encoding and articulation (Levelt et al., 1999). To accommodate the multi-word frequency effects reported here, such models would have to either attribute the articulatory effects to post-lexical processes, and then have to explain how they also happen in recognition, or allow for the selection and planning of units larger than words. In contrast, sequence-frequency effects are compatible with connectionist models of production (Chang, 2002; Chang, Dell, & Bock, 2006; Dell, Chang, & Griffin, 1999), and follow naturally from the emphasis such models place on the importance of sequential information and on the lack of distinction between 'lexicon' and 'grammar'. The challenge for such models is of a practical nature: to account for sequence-frequency effects, existing implementations would have to allow for the activation of multi-word lemmas, and allow for competition between multi-word sequences.

5.2 Modeling multi-word frequency effects: storage versus computation

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 and event structures. Sentence comprehension is affected by a multitude of distributional factors, including the frequency of words (Rayner & Duffy, 1988); the frequency of words in specific syntactic structures (verb-subcategorization biases, Clifton, Frazier, & Connine, 1984; Garnsey, Pearlmutter, Myers, & Lotocky, 1997; MacDonald, Pearlmutter, & Seidenberg, 1994); 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 versus reduced relative, Frazier & Fodor, 1978) and the event in question (Elman, 2009). Production is also affected by the distributional properties of units of varying sizes. It is affected by word frequency (Jescheniak & Levelt, 1994), by the likelihood of a word given the previous one (Jurafsky et al., 2001), as well as by the likelihood of the syntactic structure the word is part of (Gahl & Garnsey, 2004; Jaeger, 2006; Tily et al., 2009). The current findings join previous ones in adding multi-word statistics to the information impacting both production and recognition.

Finding that language users are sensitive to detailed distributional information on many levels of linguistic analysis raises questions about how to model such effects and re-opens the question of balancing storage and computation (Baayen, 2007). There are several ways one could model the multi-word effects reported here. One possibility is that the sequences have a whole-form representation, like the one used to account for frequency effects for regularly inflected words (Baayen, Dijkstra, & Schreuder, 1997). This solution minimizes computation and maximizes storage, but seems unlikely given the vast number of multi-word representations that would be required to fully model speakers' knowledge of language (Baayen et al., this volume)

Another possibility is modeling multi-word frequency effects without the presence of a whole-form representation by repeatedly calculating co-occurrence patterns, a solution that requires less storage, but more computation compared to storing the sequence. In several recent papers, Baayen and colleagues offer a way to model multi-word effects without storing multi-word sequences, or even words (Baayen & Hendrix, 2011; Baayen et al., 2011; this volume). They use a discriminative learning framework to model several processing effects, among them the multi-word frequency effects found by Arnon and Snider (2010). Using one- and two-letter sequences as input nodes, the model learns to associate letter sequences with word meanings. These associations can then be used to predict word and word sequence recognition times. The proposed model offers an elegant solution to the need to balance storage and computation, but relies on representations (letters) and computational algorithms whose psychological reality for human speakers in general, and human learners in particular, needs to be examined.

Another way to balance storage and computation is to distinguish between the building blocks used in the initial learning stages—which may well include whole-form representations of both inflected words and multi-word sequences—and the ones utilized once the relevant linguistic units have been discovered. Because infants do not know where word boundaries are (or what inflections are), their early inventory may include multi-word chunks (e.g., *I-don't-know*) that become fully analyzed later on (Abbot-Smith & Tomasello, 2006). Multi-word chunks are suggested to play an important role in learning grammar (Arnon, 2010; Arnon & Ramscar, 2012), and have been used as building blocks in several recent computational models (Beekhuizen, Bod, & Zuidema, 2013; Bod, 2009; McCauley & Christiansen, 2011).

CAPPUCINO (McCauley & Christiansen, 2011) learns language form child-directed speech by forming a chunkatory—an inventory of building blocks of varying sizes, including multi-word ones. The model has impressive cross-linguistic coverage and shows better learning than a model that does not make use of multi-word units (McCauley & Christiansen, 2011). Beekhuizen et al. (this volume) present several computational models (using Data Oriented Parsing) that use Bayesian model merging. These models allow for redundant storage and are based on the assumption that starting from unsegmented wholes is both efficient and cognitively plausible. Which solution is ultimately preferred depends on (a) the degree to which speakers show signs of holistic storage at different stages of learning (we do not view the current frequency effects as providing such evidence, but there is growing evidence for such holistic storage in the early stages of children's language learning (e.g., Bannard & Matthews, 2008; Kirjavainen & Theakston, & Lieven, 2009; Lieven et al., 2009)), (b) the possible differences in how storage and computation are weighted in humans and computers, and (c) the ability to capture both surface-level distributions and higher order properties—should such properties emerge as having a significant effect on processing.

6 Conclusion

In this paper, we find an effect of multi-word frequency on phonetic duration: durations are reduced in higher frequency sequences, for both constituents and non-constituents. The study document such effects in spontaneous speech, and is the first to examine their possible modulation by constituency. The findings show that speakers make use of multi-word statistics in production; illustrate parallels between production and comprehension; and open up a novel set of questions about the effect—or lack thereof—of higher order properties on the use of multi-word information.

Acknowledgements

We thank the reviewers and editors for their helpful comments and suggestions on this manuscript. Both authors contributed equally to this paper. We thank Sam Miller and Masoud Jasbi for their help in collecting the experimental data.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. Finding frequency effects for multi-word sequences does not mean that these sequences are represented as holistic units (for that we would need to know whether they have internal structure), but only that speakers are sensitive to their distributional information. This sensitivity can be derived without storing the sequences as independent units (see Baayen et al., 2011; this volume).
2. Some of the frequency factors that were significant in the model are not significant in model comparisons. This is probably because the comparison is done over the unresidualized predictors, which affects predictors that are highly collinear with others, such as the part frequency ones.

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.

Arnon, I. (2010). *Starting big: The role of multi-word phrases in language learning and use*. Unpublished PhD dissertation, Stanford University.

Arnon, I., & Clark, E. V. (2011). When '*on your feet*' is better than '*feet*': Children's word production is facilitated in familiar sentence-frames. *Language Learning and Development*, 7, 107–129.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: how order-of-acquisition affects what gets learned. *Cognition*, *122*, 292–305.

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

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for the relationship between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*, 31–56.

Baayen, R. H. (2007). Storage and computation in the mental lexicon. In G. Jaremaand, & G. Libben (Eds.), *The Mental Lexicon: Core Perspectives* (pp. 81–104). Elsevier.

Baayen, R. H., Dijkstra, A., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language, 37*, 94–117.

Baayen, R. H., & Hendrix, P. (2011). Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. LSA workshop 'Empirically examining parsimony and redundancy in usage-based models'.

Baayen, R. H., Milin, P., Filipovic Durdevic, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*, 438–481.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science*, *19*, 241–248.

Bard, E. G., Anderson, A., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, *42*, 1–22.

Beekhuizen, B., Bod, R., & Zuidema, W. (2013). Three design principles of language: the search for parsiin redundancy.*Language and speech, 56*.

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. *Journal of the Acoustical Society of America*, *113*, 1001–1024.

Bod, R. (2009) From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science*, *33*, 752–793.

Bybee, J. L. (1998). The emergent lexicon. In M. C. Gruber, D. Higgins, K. S. Olson, & T. Wysocki (Eds.), *Proceedings of the Chicago Linguistic Society 34: The Panels* (pp. 421–435). Chicago, IL: Chicago Linguistic Society.

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. (2002) Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, *26*, 609–651.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*, 234–272.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Cieri, C., Miller, D., & Walker, K. (2004). The Fisher Corpus: A Resource for the Next Generations of Speech-to-Text. In: Proceedings of the Language Resources and Evaluation Conference.

Clifton, C., Frazier, L., & Connine, C. (1984). Lexical expectations in sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, *23*, 696–708.

Dell, G. S., Chang, F., & Griffin, Z. M. (1999). Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science*, *23*, 517–542.

Deshmukh, N., Ganapathiraju, A., Gleeson, A., Hamaker, J., & Picone, J. (1998). Resegmentation of the SWITCHBOARD. In Fifth International Conference on Spoken Language Processing.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, *41*, 375–396.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, *33*, 547–582.

Frank, A., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. *The 30th Annual Meeting of the Cognitive Science Society (CogSci08)*, pp. 939–944.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*, 829–834.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *66*, 291–325.

Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, *80*, 748–775.

Garnsey, S., Pearlmutter, N., Myers, E., & 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.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. Proceedings of ICASSP–*92*, 517–520.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. London: Oxford University Press.

Jaeger, T. F. (2006). *Redundancy and syntactic reduction in spontaneous speech*. Unpublished Dissertation, Stanford University, Stanford, CA.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*, 23–62.

Jannsen, N., & Barber, H. A. (2012). Phrase frequency effects in production. *PLoS ONE, 7(3): e33202. doi:10.1371/journal.pone.0033202.*

Jescheniak, J., & Levelt, W. J. M. (1994). Word frequency effects in production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 824–843.

Jurafsky, D., Bell, A., Gregory, M. L., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. L. Bybee, & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 229–254). Amsterdam, the Netherlands: Benjamins.

Kirjavainen, M. M. M., Theakston, A. L., & Lieven, E. V. (2009). Can input explain children's me-for-I errors? *Journal of Child Language*, *36*, 1091–1114.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.

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) (Vol. 19, pp. 849–856). Cambridge, MA: MIT Press.

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children’s production of multiword utterances: A usage-based analysis. *Cognitive Linguistics, 20*, 481–507.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676–703.

Marantz, A. P. (1981). *On the nature of grammatical relations*. PhD thesis, Massachusetts Institute of Technology.

McCauley, S. M., & Christiansen, M. H. (2011). Learning simple statistics for language comprehension and production: The CAPPUCCINO model. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 1619–1624). Austin, TX: Cognitive Science Society.

McClelland, J. L. (2010). Emergence in cognitive science. *Topics in Cognitive Science, 2*, 751–770.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to understanding cognition. *Trends in Cognitive Sciences*, *14*, 348–356.

Miller, J., & Weinert, R. (1998). *Spontaneous Spoken Language: Syntax and Discourse*. Oxford: Clarendon Press.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Harper Collins.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*, 456–463.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*, 191–201.

Reali, F., & Christiansen, M. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*, *57*, 1–23.

Sag, I. A. (2013). Sign-based construction grammar: An informal synopsis. To appear in H. C. Boas & I. A. Sag (Eds.), *Sign-based construction grammar*. Stanford, CA: CSLI Publications.

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. (2011). Seeing a phrase ‘time and again’ matters: The role of phrasal frequency in the processing of multi-word sequences. *Journal of Experimental Psychology: Language, Memory, and Cognition, 37*, 776–784.

Smith, M., & Wheeldon, L. (1999). High level processing scope in spoken sentence production. *Cognition*, *73*, 205–246.

Snider, N., & Arnon, I. (2012). A unified lexicon and grammar? Compositional and non-compositional phrases in the lexicon. In S. Gries & D. Divjak (Eds.), *Frequency effects in language*. Berlin, Germany: Mouton de Gruyter.

Tily, H., Gahl, S., Arnon, I., Kothari, A., Snider, N., & Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language & Cognition*, *1*, 147–165.

Tremblay, A., & Baayen, H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on formulaic language; Acquisition and communication* (pp. 151–173). London: The Continuum International Publishing Group.

Tremblay, A., Derwing, B., Libben, G., & Westbury, G. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, *61*, 569–613.

Tremblay, A., & Tucker, B. V. (2011). The effect of N-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon*, *6*, 302–324.

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., & Walenski, M. (2005). Reply: Moving past the past tense. Brain and Language, *93*, 248–252.

Weide, R. (1998). *The CMU Pronunciation Dictionary*, release 0.6. Carnegie Mellon University.