Determinants of Lengths of Repetition Disfluencies: Probabilistic syntactic constituency in speech production

Zara Harmon and Vsevolod Kapatsinski
University of Oregon

1 Introduction

Usage-based theories of grammar suggest that constituent structure emerges in part from co-occurrence: items used together fuse together forming cohesive, hard-to-interrupt units (Bybee 2002, see also Gregory et al. 1999, Kapatsinski 2010, Stefanowitsch & Gries 2003). This study is an effort to investigate the effects of co-occurrence on constituent structure in language production. We investigate these effects by looking at repetition disfluencies, in which one or more elements in the sentence are repeated after an interruption point in speech. Repetition disfluencies have been argued to be sensititive to constituency: the speaker restarts production from the major constituent boundary nearest to the point at which the flow of speech was interrupted (the "interruption point", Clark & Wasow 1998, DuBois 1974, Fox & Jasperson 1995, Levelt 1989). This follows from the more general hypothesis that the more cohesive a unit, the less likely it is to be interrupted (Kapatsinski 2010). For example, speech production is never restarted from the middle of a word. Given the proposed influence of constituency on repair, we suspected that probabilistic influences on constituency may also affect repetition repair, by making syntactic constituent boundaries located between frequently co-occurring words weaker and thus making the speaker less likely to restart production from those weakened boundaries.

1.1 The structure of repetition repairs

A repetition repair involves an interruption in the flow of speech when the speaker decides to repair part of the already produced utterance or repeat something s/he just said to buy time to plan upcoming speech (Fox & Jasperson 1995, Kapatsinski 2010). The form of such a disfluency contains an interruption point which is followed by a recycling of one or more words, as in (1)-(3) (data from Switchboard Corpus, Godfrey *et al.* 1992). We will refer to one-word repetitions as in (1) unigrams, in (2) bigrams, in (3) trigrams, and to all repetitions as n-grams.

(1) I really appreciated [the, + **the**] whole, uh, English class
(2) The crime level is not as high as it is in other areas [of the, + **of the**] city
(3) I [had a similar, + **had a similar**] health plan

We call the first word before the interruption point "word0", the second word "word1", and the third word "word2". We will refer to the boundary between word1 and word0 as transition [10]; the boundary between word2 and word1 as

transition [21]; and the boundary between word3 and word2 as transition [32]. This is illustrated in Figure 1.

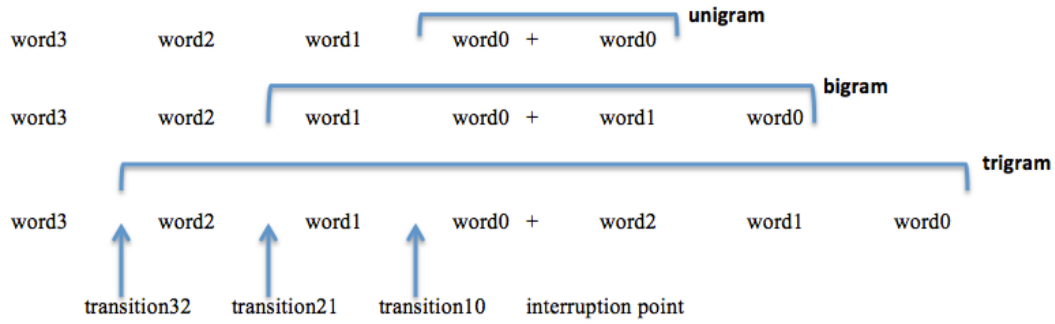

Figure 1: The Structure of a Disfluent Event

As noted above, though interruptions sometimes occur in the middle of a word one never restarts the recycle from the middle of a word, hence the ungrammaticality of cases in (5) and (6), cf. (4).

(4) Why should they be [sit-, + **sitting**] in, uh, prison getting their college degree
(5) *I [had a similar, + **-milar**] health plan
(6) *Why should they be [sit-, + **-ting**] in, uh, prison getting their college degree

This robust observation provides strong evidence for the hypothesis that restarts are sensitive to constituency: one does not restart from the middle of a cohesive unit. In this paper, we ask whether the same restriction would apply to units larger than single words. In accordance with usage-based approaches to constituency, we hypothesized that cohesion of a word n-gram is influenced by the co-occurrence statistics of the words forming the n-gram.

These questions were previously examined in a small-scale investigation (Kapatsinski 2005), which suggested that speech after the interruption point was restarted from the nearest major syntactic constituent boundary (Levelt 1989), *unless* that boundary was high in backward transitional probability. Backward transitional probability is the probability of a word given the following word. For example, in sentence (1), the recycle starts from the transition before *the* i.e transition [01] resulting in a unigram disfluency. We can account for this in at least four ways. First, we could argue that transition [01] is a Direct Object boundary, which is a major syntactic constituent boundary and attracts restarts. Second, we could argue that by default, speakers' preference is to avoid repeating more than one word. Third, if the speaker generally tends to repeat as little as possible, the word *appreciated* may not be repeated because it is long. Finally, we could argue that backward transitional probability accounts for the place of

restart: The probability of *appreciated* given that the following word is *the* is low, and so *appreciated the* is not a cohesive unit and, as a result, prone to interruption.

The present study follows Kapatsinski (2005) in arguing for the fourth explanation. Compared to that study, we greatly expanded the size of the sample to 2500 instances of repetition repairs. We also coded word length (in phonemes) to take into account the fact that frequent and probable words tend to be shorter (e.g. Zipf 1949); given that frequent words are shorter, infrequent words may be less likely to be repeated if the speaker repeats as little as possible. Finally, we examined an additional sample in which syntactic structure was strictly controlled. This is important because probabilistic measures are correlated with syntax, and backwards transitional probability, the strongest probabilistic predictor in Kapatsinski (2005), has the strongest correlation with syntax of all in preposing languages like English.

2 Methods

The data was retrieved from the Switchboard Corpus of telephone conversations (Godfrey et al. 1992). Our main sample was created by randomly sampling all instances of repetition of a certain length meeting our inclusion criteria. The data comprised 1000 instances of unigrams selected by random sampling from 10565 (48.03%) instances; 1000 instances of bigrams selected by random sampling from 2915 (86.76%) instances; and finally all 500 instances of trigrams. As evident from these numbers, speakers predominantly repeat single words.

Exclusions were mainly cases in which the interruption point followed the first word of the clause or, for the purposes of comparing bigrams to trigrams, the second word. These cases were excluded because recycles are never strings crossing clause boundaries. We also excluded complex disfluencies, in which the repetition is immediately preceded or followed by anther disfluency. Finally, abandonments, in which the recycle was abandoned and the continuation was never produced, were also excluded.

2.1 Probabilistic Measures

We used 5 local measures of probabilistic relations between words. These include word frequency (absolute), string frequency, forward and backward transitional probability, and mutual information.

Word frequency is the absolute token frequency of a word (total number of occurrence) in the corpus. Of all the probabilistic measures used in this study, word frequency is the only measure that does not take into account any association between a word and its neighboring words and so it cannot measure cohesiveness of a multi-word unit. It is included here because frequent words are semantically and phonologically ‘lighter’ and may be more repeatable for these reasons alone. String frequency (Krug 1998), or joint probability of a two-word string, is the frequency of the string in the corpus.

Forward transitional probability refers to the conditional probability of a word

given the preceding word (Jurafsky et al. 2001). Forward transitional probability of a two-word string is the frequency of the string divided by the frequency of the first word in the string. Forward transitional probability can be taken as a measure of predictability during 'normal', left-to-right planning of an utterance (e.g. Bybee 2001:163-164, Elman 1990).

Backward transitional probability refers to the conditional probability of a word given the following word. Backward transitional probability of a two-word string is the frequency of the string divided by the frequency of the second word in the string. Backward transitional probability can be thought of as a measure of predictability during backtracking, as the speaker goes back through the string to find the nearest constituent boundary (Kapatsinski 2005) or predicts a preceding word from the following word (e.g. when the preceding word was obscured by acoustic or perceptual noise). It is also the probabilistic measure best correlated with syntactic constituency in English and other preposing languages. It can therefore be seen as an index of syntactic constituency rather than an independent influence.

Mutual information of a two-word string is the frequency of the string divided by the product of the frequencies of the two words forming the string. Mutual information differs from backward and forward transitional probability in that it is not directional (e.g. Gregory *et al.* 1999, Gries 2013). Both mutual information and string frequency have been used as overall indicators of association between two words, with mutual information controlling for how often the two words would be expected to occur by chance. Mutual information can thus be thought of as an outcome of Hebbian learning, where an association between two words is increased when they occur together and weakened when one of the words occurs without the other (Hebb 1949). However, it has the possible disadvantage of weighting both forward and backward processing directions equally (e.g. Bybee 2001:163, Gries 2013).

In the first part of the study, we will look at these probabilistic factors to determine which ones are promising in predicting the length of the repetition. We will present four analyses: in the first analysis we will attempt to predict the choice between unigrams versus bigrams. In the second analysis we will model the choice between bigrams and trigrams. The third and fourth analyses will look at whether these probabilistic measures have any predictable power beyond syntax and word length.

2.2 Phonological and Syntactic Coding

To investigate the role of phonological factors in predicting the location of restart, we measured the lengths of each of the four words preceding the interruption point. The length was measured in phonemes based on trascriptions from the Carnegie Mellon University Pronouncing Dictionary (Weide 1998).

The syntactic category of each of the four words preceding the interruption point was also coded by hand. The categories were coded as narrowly as possible, allowing the statistical model to discover any possible groupings that exist. This

was done to ensure that syntax is allowed to explain as much variance as it can, which in turn ensures that any probabilistic effects detected are not explainable by syntactic structure. Interactions between syntactic categories of the various words in the string preceding the interruption point were also allowed, so that effects of various types of syntactic boundaries could be detected.

2.3 Statistical Analysis

Variable importance measures derived from random forests of conditional inference trees (Strobl *et al.* 2009) were used to find the most important predictors of the length of the disfluency. All statistical analyses were conducted in R using (Party) package. Variable importance provides information on how important it is to include a predictor in the tree to achieve a good model. There are a few advantages to using these statistical methods. First, random forests facilitate analyzing correlated predictors (Strobl *et al.* 2008). Also unlike regression, conditional inference trees can tackle the problem of unattested combinations of predictor values, which is a common occurrence in corpus studies and is particularly true for this study, since many sequences of syntactic categories are ungrammatical and therefore never occur.

3 Results

3.1 Analysis I: Probabilistic predictors of repeating more than one word

In this analysis, we asked why the speaker would go back further than one word. So, the dependent variable in this analysis was binary: unigrams versus longer repetitions. We consider three possible explanations for repeating more than one word. It is possible that the speaker skips over frequent words. In this case, we expect to see token frequency of word1 as an important predictor. It is also possible that the speaker is skipping over cohesive transitions. In this case we expect to see one of the measures of cohesion (string frequency, transitional probability or mutual information) of transition [10] as an important predictor. The final possible explanation is attraction towards far-away low-cohesion transitions. In this case, we expect to see measures of cohesion of transitions [21] and/or [32] as important predictors.

As illustrated in Figure 2, the best probabilistic predictor of whether one would skip over transition [10] is backward transitional probability of transition [10]. Frequencies of word0 and word1 as well as string frequency and mutual information of transition [10] appear to be not as predictive as backward transitional probability. Characteristics of further-away transitions [21] and [32] do not appear to be predictive of whether the speaker restarts from those transitions.

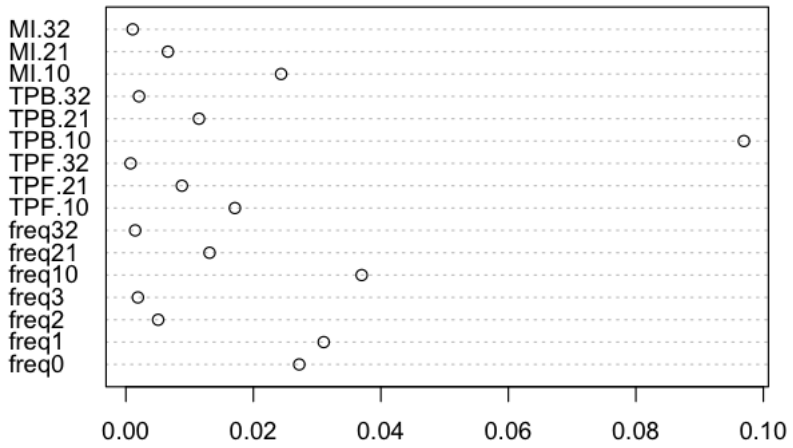

Figure 2: Variable importance measures for predicting whether one skips over the nearest word boundary (a.k.a. 10). TPB is backward transitional probability; TPF is forward transitional probability; MI is mutual information; freq is frequency. Random forest variable importance measures are between 0 and 0.5, with 0 being least predictive.

Thus, it is the characteristics of *radio in* and *on a* in (7) and (8) that appear to be most predictive of the fact that the speaker only repeats one word in (7) and two words in (8). In (7), the backward transitional probability of *radio* given *in* is low and so transition [10] is a low-cohesion transition and prone to interruption. However in (8), the backward transitional probability of *on* given *a* is high and so transition [10] is highly cohesive transition and therefore less interruptible, which is why the speaker skips back to transition [21] to start the repetition from. A smaller contribution is also likely made by the fact that *radio* is less frequent than *on*.

(7) ...have the national public radio [in + in] my area
(8) ...an English translation [on a + on a] screen

3.2 Analysis II: Probabilistic predictors of repeating two vs. three words

In this analysis, we compared bigrams to trigrams. This analysis asks the following question: if the speaker has not stopped at the first transition, why would they stop at the second (transition [21]) versus the third (transition [32])? Again, there are three possible explanations for this. It is possible that the speaker skips over frequent words. In this case, we expect to see word frequency of word2 as an important predictor. It is also possible that the speaker is skipping over cohesive transitions. In this case we expect to see one of the measures of cohesion of transition [21] as an important predictor. The final possible explanation is attraction towards far-away low-cohesion transitions. In this case, we expect to see one of the measures of cohesion of transition [32] as important predictors.

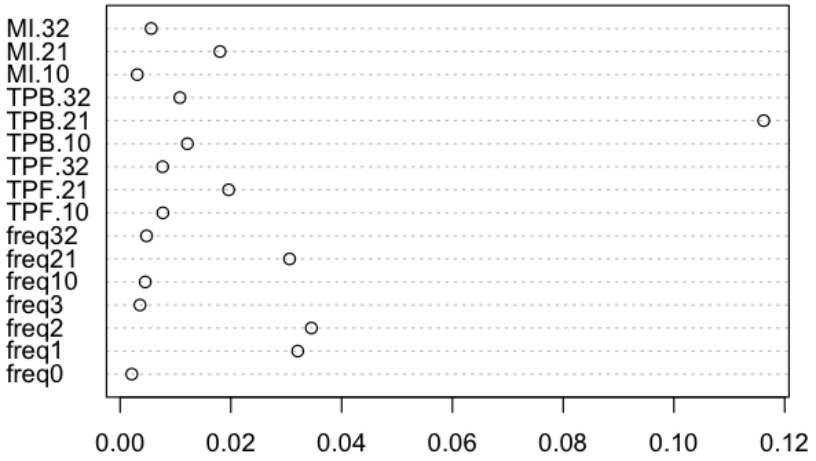

Figure 3: Variable importance measures for predicting whether one stops at transition [21]

As illustrated in Figure 3, the results indicate that the best probabilistic predictor of why one would stop at the second transition (transition [21]) is backward transitional probability of transition [21]. Frequency of word1 and word2 as well as string frequency of transition [21] are also strong predictors but not as strong as backward transitional probability. For example, comparing disfluencies in (9) and (10) below, the characteristics of *translation on* and *lot of* and not the characteristics of the further back transitions (*English translation* or *a lot*) that best predict the lengths of the recycles in those examples. In (9), the backward transitional probability of *translation* given *on* is low and so transition [21] is a low-cohesion transition and prone to interruption. However in (10), the backward transitional probability of *lot* given *of* is high and so transition [21] is a highly cohesive transition and less interruptible, which is why the speaker skips back to transition [32] to start the repetition.

(9) ...an English translation [on a + on a] screen
(10) ... and a [lot of the + lot of the] songs have some lyrics

So far, we have observed that the best probabilistic predictor of how far back the speaker would go to start a repetition is backward transitional probability. In other words, the speaker skips over cohesive transitions until s/he gets to a less cohesive transition and that transition would be where the repetition starts from. It is also likely that the speaker is skipping over frequent words. However, backwards transitional probability is highly correlated with syntax and word frequency is negatively correlated with word length. Therefore, we conducted the following two analyses to determine whether probabilistic influences on constituency help predict recycle length controlling for syntax and word length.

3.3 Analysis III: The importance of syntax and phonology

When syntax is entered into the analysis (Figure 4), it appears to play a dominant role in predicting the length of disfluency. In the presence of syntax as a predictor, backward transitional probability does not appear as important as it has in Analyses I-II, suggesting that it's apparent importance in Analyses I and II may be due to its high correlation with syntactic constituency. The length of a word appears to be of at most secondary importance in predicting whether the word is repeated. Rather, the best predictor of whether the word is skipped over appears to be the syntactic category of that word. Thus, whether one skips over transition [10] is predicted by the syntactic category of word1 (Figure 4, left), and whether one skips transition [21] is predicted by the syntactic category of word2 (Figure 4, right). Given the importance of syntax, Analysis IV reports on a set of disfluencies in which the syntactic structure of the preceding three-word string is controlled.

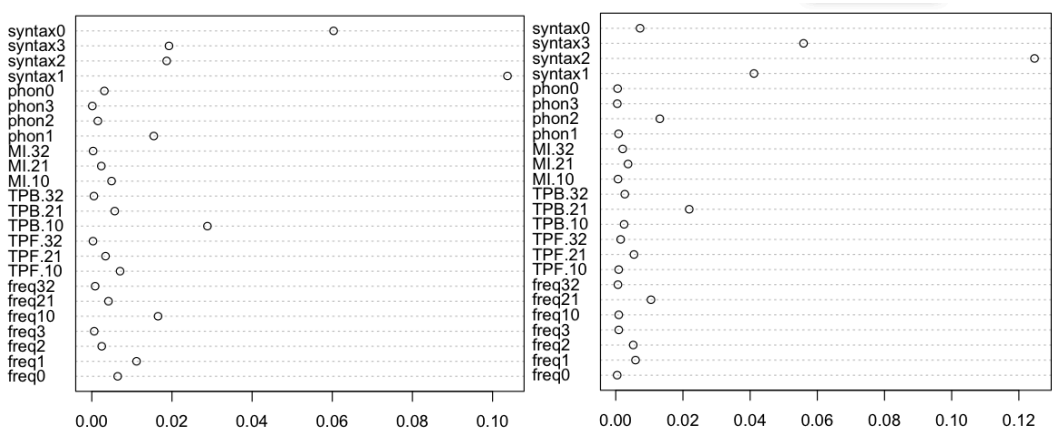

Figure 4: The importance of syntax for predicting whether one skips over the nearest word boundary (transition [10], left) and for predicting whether one stops at transition [21] (right). Phon stands for word length in phonemes; syntax stands for the syntactic categories of words.

3.4 Analysis IV: Probabilistic predictors in the presence of syntax

The sample of disfluencies in this analysis consisted of repetitions in which the interruption point was immediately preceded by a three-word string consisting of verb as word2, a preposition as word1, and a determiner or a noun as word0. In this sample, the syntactic categories of the potentially skipped-over words were controlled: word1 was always a preposition, and word2 was always a verb. The sample consisted of 301 instances, with 248 unigram, 49 bigram and 3 trigram repetitions. Because of the low number of trigram repetitions, we attempted to predict only whether the speaker repeated more than one word.

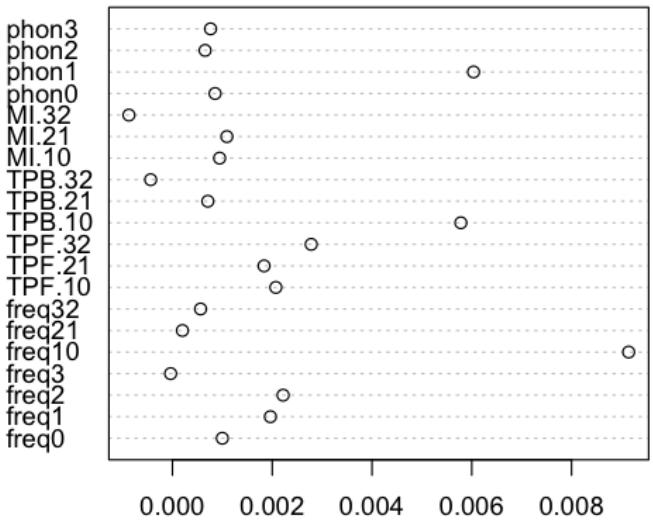

Figure 5: The importance of probabilistic, syntactic, and phonological factors for predicting whether one skips over the nearest word boundary (transition [10]) where word1 is a preposition and word2 is a verb.

The results are shown in Figure 5. In this highly controlled sample, the most important predictor of whether one skips over transition [10], repeating more than one word, is the frequency of the two-word string spanning transition [10]. As in previous analyses, characteristics of transitions [21] and [32] are unimportant. Even though the length of the word that is potentially skipped over (word1) is also important, it is not as predictive as string frequency. These results suggest that frequency of co-occurrence influences interruptibility of a word string when syntax is controlled, even in the presence of phonological predictors.

Figure 6 illustrates the effects of string frequency and backward transitional probability by plotting the most probable tree with the important predictors from the random forest entered into the model. As the figure shows, the effects of frequency and probability are in the expected direction: frequent strings and probable words are likely to be skipped over. When frequency is removed from the analysis, the effect of word length emerges, also in the expected direction: shorter words are more likely to be repeated than longer words. [1]

[1] We have also analyzed these data using logistic regression (with frequency and transitional probability rank-transformed to reduce skew). The directions of all effects remain unchanged. String frequency remains a significant predictor (z = 2.01, p = .04); length of word1 is also significant (z = -2.14, p = .03); transitional probability is not significant.

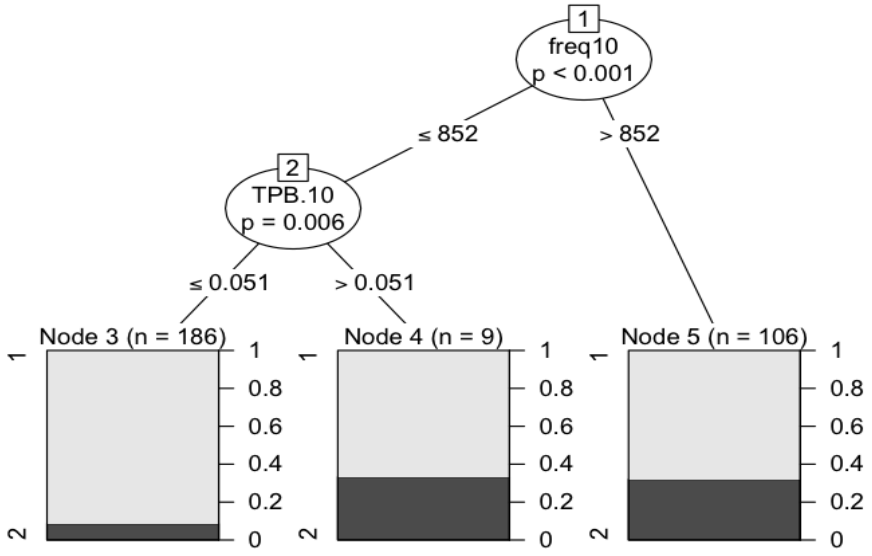

Figure 6: The most likely conditional inference tree with string frequency and backward transitional probability for transition [10], and length of word1 as predictors. The dark sections of the bins represent the probability of repeating two or more words. The light sections represent probability of repeating a single word.

4 Discussion

The importance of co-occurrence frequency in predicting recycle length when syntax is controlled attests to the existence of probabilistic syntagmatic relations between words in the mind of the speaker. The results also show that cohesion is probabilistic, and not just categorical: When syntax is controlled, there are still effects of co-occurrence on interruptibility of word strings. In turn, the finding that there is interaction between probabilistic biases and syntactic constituency, provides evidence for constituency as a probabilistic tendency.

Our findings contribute to a larger literature on the influence of probabilistic factors on cohesion. For example, Khakimov (2014, GURT) shows that codeswitches also seem to occur at boundaries low in backward transitional probability. In a study on replacement disfluencies, Kapatsinski (2010) shows that frequent chunks are more likely to be completed before being replaced. The same effect has been found in stop-signal experiments (Logan 1982). Furthermore, letters and sounds are hard to detect in cohesive chunks (Goldman-Eisler 1968, Healy 1976, Sosa & MacFarlane 2002, Kapatsinski & Radicke 2009). Several studies also point to the influence of probabilistic prediction in speech perception and reading (Seidenberg & MacDonald 1999, Dahan & Tanenhaus 2005, Levy 2008, Reali & Christiansen 2007).

We considered 5 probabilistic predictors: word frequency, string frequency, forward and backward transitional probability, and mutual information. The results showed that, on its own, backwards transitional probability is the best

predictor of the length of the recycle (replicating Kapatsinski 2005). However, backward transitional probability is highly correlated with syntactic constituency, and syntax is shown to be the stronger predictor of interruptibility (Analysis III). Once syntax is controlled (Analsysis IV), the importance of transitional probability diminishes, suggesting that the greater predictiveness of backwards transitional probability is due, at least in part, to its correlation with syntactic constituency. String frequency (a measure of chunking rather than predictability, cf. Bybee 2001:161-165, Gregory *et al.* 1999, Jurafsky *et al.* 2001) emerges as the best predictor of interruptibility.

In all analyses reported, the speaker tends to avoid repeating more than a single word, and repetitions of more than two words are exceedingly rare. Furthermore, whether the speaker goes further back than transition *n* depends on characteristics of transition *n*: cohesive transitions repel restarts further back rather than low-cohesion transitions further back attracting them. One account of these results is that the speaker tends to recycle a single chunk. A single chunk is usually a single word but can be longer when that word is part of a frequent multiword string / collocation. A possible exception to this generalization is presented by clause boundaries: it appears that the clause boundary can actively attract restarts; perhaps, by allowing the listener to revise their interpretation of the clause more easily or because the clause is itself a higher-level chunk or schema in production that the speaker is driven to restart from the beginning.

References

Bybee, J. 2001. *Phonology and Language Use*. Cambridge: Cambridge University Press.

Bybee, J. 2002. Sequentiality as the basis of constituent structure. In: *The Evolution of Language out of Pre-Language*, ed. By T. Givon & B. F. Malle, 109-134. Amsterdam: John Benjamins.

Clark, H. H., & T. Wasow. 1998. Repeating words in spontaneous speech. *Cognitive Psychology* 37.201-242.

Dahan, D., & M. K. Tanenhaus. 2005. Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin & Review* 12.453-459.

DuBois, John W. 1974. Syntax in mid-sentence. In *Berkeley Studies in Syntax and Semantics I*, ed. By C. Fillmore, G. Lakoff & R. Lakoff, pp. III.1-25. Berkeley, CA: Department of Linguistics and Institute of Human Learning, University of California.

Elman, J. L. 1990. Finding structure in time. *Cognitive Science* 14.179-211.

Fox, B. A. & R. Jasperson. 1995. A syntactic exploration of repair in English conversation. In *Alternative Linguistics: Descriptive and Theoretical Modes*, ed. By P. W. Davis, 77-134. Amsterdam: John Benjamins.

Godfrey, J. J., E. C. Holliman, & J. McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. *IEEE ICASSP*, I517-I520.

Goldman-Eisler, F. 1968. *Psycholinguistics: Experiments in Spontaneous Speech*. London: Academic Press.

Gregory, M. L., W. D. Raymond, A. Bell, E. Fosler-Lussier, & D. Jurafsky. 1999. The effects of collocational strength and contextual predictability in lexical production. *Chicago Linguistic Society* 35.151-166.

Gries, S. Th. 2013. 50-something years of work on collocations: what is or should be next… *International Journal of Corpus Linguistics* 18. 137-166.
Healey, A. F. 1976. Detection errors on the word *the*: Evidence for reading units larger than letters. *Journal of Experimental Psychology: Human Perception & Performance* 6.403-409.
Hebb, D. O. 2002. *The Organization of Behavior: A Neuropsychological Theory*. New York: Psychology Press.
Jurafsky, D., A. Bell, M. Gregory, & W. D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In *Frequency and the Emergence of Linguistic Structure*, ed. By J. L. Bybee & P. J. Hopper, 229-254. Amsterdam: John Benjamins.
Kapatsinski, V. 2010. Frequency of use leads to automaticity of production: Evidence from repair in conversation. *Language & Speech* 53.71-105.
Kapatsinski, V. 2005. Measuring the relationship of structure to use: Determinants of the extent of recycle in repetition repair. *Berkeley Linguistics Society* 30.481–492.
Kapatsinski, V., & Radicke, J. 2009. Frequency and the emergence of prefabs: Evidence from monitoring. In *Formulaic Language: Volume 2. Acquisition, Loss, Psychological Reality, and Functional Explanations*, ed. By R. Corrigan, E. A. Moravcsik, H. Ouali & K. Wheatley, 499-520. Amsterdam: John Benjamins.
Khakimov, N. 2014. A usage-based approach to Russian-German code-mixing. Paper presented at Georgetown University Round Table of Languages and Linguistics GURT. March 14-16, 2014. Georgetown University, Washington, D. C.
Krug, M. 1998. String frequency: A cognitive motivating factor in coalescence, language processing and linguistic change. *Journal of English Linguistics* 26.286-320.
Levelt, W. J. M. 1989. *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.
Levy, R. 2008. Expectation-based syntactic comprehension. *Cognition* 106.1126-1177.
Logan, G. D. 1982. On the ability to inhibit complex movements: A stop-signal study of typewriting. *Journal of Experimental Psychology: Human Perception and Performance* 8.778-792.
Reali, F., & M. H. Christiansen. 2007. Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory & Language* 57.1-23.
Seidenberg, M. S., & M. C. MacDonald. 1999. A probabilistic constraints approach to language acquisition and processing. *Cognitive Science* 23.569-588.
Shriberg, E. 1999. Phonetic consequences of speech disfluency. *Proc. 14th ICPhS, San Francisco,* 619-622.
Sosa, A. V., & J. MacFarlane. 2002. Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of*. *Brain & Language* 83.227–236.
Stefanowitsch, A., & S. Th. Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8.209–243.
Strobl, C., A. L. Boulesteix, T. Kneib, T. Augustin, & A. Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9.307.
Strobl, C., J. Malley, & T. Gerhard. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14.323–348.
Weide, R. 1998. CMU Pronouncing Dictionary. http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/cmudict.0.7a
Zipf, G. K. 1949. *Human Behavior and the Principle of Least Effort*. Oxford: Addison-Wesley.