Effects of Duration, Locality, and Surprisal in Speech Disfluency Prediction in English Spontaneous Speech

Samvit Dammalapati
IIT Delhi
samvit1998@gmail.com

Rajakrishnan Rajkumar
IISER Bhopal
rajak@iiserb.ac.in

Sumeet Agarwal
IIT Delhi
sumeet@iitd.ac.in

Abstract

This study examines the role of two influential theories of language processing, Surprisal Theory and Dependency Locality Theory (DLT), in predicting disfluencies (fillers and reparandums) in the Switchboard corpus of English conversational speech. Using Generalized Linear Mixed Models for this task, we incorporate syntactic factors (DLT-inspired costs and syntactic surprisal) in addition to lexical surprisal and duration, thus going beyond the local lexical frequency and predictability used in previous work on modelling word durations in Switchboard speech (Bell et al., 2003, 2009). Our results indicate that compared to fluent words, words preceding disfluencies tend to have lower lexical surprisal (hence higher activation levels) and lower syntactic complexity (low DLT costs and low syntactic surprisal except for reparandums). Disfluencies tend to occur before upcoming difficulties, *i.e.,* high lexical surprisal words (low activation levels) with high syntactic complexity (high DLT costs and high syntactic surprisal). Further, we see that reparandums behave almost similarly to disfluent fillers with differences possibly arising due to effects being present in the word choice of the reparandum, *i.e.*, in the disfluency itself rather than surrounding it. Moreover, words preceding disfluencies tend to be function words and have longer durations compared to their fluent counterparts, and word duration is a very effective predictor of disfluencies. Overall, speakers may be leveraging the differences in access between content and function words during planning as part of a mechanism to adapt for disfluencies while coordinating between planning and articulation as suggested by Bell et al. (2009).

1 Introduction

One of the primary reasons for disfluencies in speech is difficulties in language production (Tree and Clark, 1997; Clark and Wasow, 1998). In this study, we investigate the impact of predictability and working memory measures of processing complexity based on two influential linguistic theories, *viz.*, Surprisal Theory (Levy, 2008; Hale, 2001) and Dependency Locality Theory (DLT Gibson, 2000), in predicting the following two types of disfluencies:

1. *Disfluent fillers*: Utterances like *uh*, *um* which break fluency by interjecting and creating an interruption between words (as in the spoken utterance "thinking about the *uh* day when I").

2. *Reparandums*: Cases where speakers make corrections in their speech. For example, when a speaker says "Go *to the righ-* to the left". Here, the speaker makes a correction to *to the righ-* (reparandum) by restarting with the intended (corrected) speech *to the left* (repair).

We situate our work in the widely accepted framework of speech production models proposed by Levelt and collaborators (Levelt, 1992; Bock and Levelt, 1994; Levelt et al., 1999) which conceive speech production as comprising of the following stages: conceptual and syntactic planning, lexical selection, morphological and phonological encoding, and articulatory execution. Previous work analyzing Switchboard speech (Bell et al., 2009, 2003) showed evidence that lexical frequency and predictability are significant predictors of word durations in spontaneous speech. Further, a long line of work notes that words occurring before disfluencies (usually function words) are lengthened by speakers (Bell et al., 2003; Tree and Clark, 1997; Shriberg, 1995). To account for this effect, Bell et al. (2009) put forth the proposal that elongation is part of a mechanism to adapt for disfluencies, while coordinating between planning and articulation. Thus disfluencies are an outcome of coor-

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 91-101.
Held on-line February 14-19, 2021

dination failure between planning and articulation processes *i.e.*, incomplete plans being fed into articulatory routines. We investigate the above proposal by incorporating syntactic factors (DLT-inspired costs and syntactic surprisal) into a disfluency prediction classifier, in addition to the lexical factors used in earlier corpus-based work cited above.

We examined whether lexical and syntactic surprisal measures, DLT integration and storage costs (measures of syntactic complexity) as well as duration were significant predictors of disfluencies in transcribed data from the Switchboard corpus of American English spontaneous speech (Godfrey et al., 1992). We incorporated the aforementioned predictors as fixed effects in Generalized Linear Mixed Models (GLMMs Pinheiro and Bates, 2000) with words as random effects to predict whether disfluencies existed at all words in an utterance (with disfluencies stripped off). Though these measures (and the underlying theories stated at the outset) were originally proposed to model language comprehension, recent works have demonstrated how they reflect language production difficulty too. Demberg et al. (2012) showed that syntactic surprisal, an information-theoretic measure of comprehension difficulty defined by Surprisal Theory, is a significant predictor of word duration in spontaneous speech even amidst competing controls like lexical frequency. More recently, Scontras et al. (2015) showed that for English relative clause production, locality considerations resulted in greater speech disfluencies and starting time for object relatives compared to subject relatives.

Our disfluency prediction results indicate that preceding word duration, following word lexical surprisal and DLT storage costs are the best predictors of both fillers and reparandums. Further, the regression coefficients revealed the preponderance of high processing costs (surprisal and DLT) on words following disfluencies and lower costs on words preceding disfluencies. Thus words ahead of disfluencies have low activation levels (high lexical surprisal) and high syntactic difficulties (high DLT costs and high syntactic surprisal) which makes their construction difficult. Conversely, words before disfluencies tend to be words with high activation levels (low lexical surprisal values) and low syntactic difficulty (low DLT costs and low syntactic surprisal except for reparandums). We also find that speakers take longer to articulate words before disfluencies. Thus, our results suggest that in incremental production, speakers choose highly activated points of low processing load (low lexical surprisal and low syntactic costs) to plan for upcoming difficulties and this process takes a fair amount of time and mental resources.

We propose that in order to maintain the temporal coordination between the articulatory stream and the planning of utterances, speakers lengthen words as a means to buy time for planning. The idea of such links between planning and articulation have also been suggested by Pierrehumbert (2002) in the form of "ease of retrieval" in phonological encoding and by Munson (2007) in relation to the longer and fuller articulation of disfluencies. We also observe that the words in the reparandum and the words surrounding disfluent fillers are less likely to be content words which we explain as speakers trying to pick easier words (function words are low in information and simpler in construction) around disfluencies in order to help the planning process. Further, our results also lend credence to the insight from the production literature that function words and content words have distinct modes of access (Garrett, 1975, 1980; Lapointe and Dell, 1989) and in the presence of disfluencies, speakers may be making use of these special modes of access function words have.

2 Background

In the context of disfluency detection, disfluent fillers tend to be easier to identify as they mostly consist of a closed set of fixed utterances (e.g. *um*, *uh*). Reparandums on the other hand are more difficult to identify because they tend to resemble fluent words a lot more. One of the effective feature types for detecting these reparandums are distance and pattern matching based features that look into the similarity of words and POS tags with their neighbours (Honnibal and Johnson, 2014; Zayats et al., 2014, 2016; Wang et al., 2017). The reason for their effectiveness could stem from how the repair that follows the reparandum is usually a "rough copy" of the reparandum, i.e., it incorporates how the repair has very similar words in roughly the same word order as the reparandum. Apart from this, disfluency detection has also been shown to be effective with other features like language models and lexical features (Zwarts and Johnson, 2011; Zayats et al., 2016); prosody (Shriberg et al., 1997; Kahn et al., 2005; Tran et al., 2017) and dependency based features (Honnibal and Johnson,

2014). Seeing how disfluency detection in the past has harnessed features based on lexical language models, dependency grammar and prosody, we examined whether disfluencies can be explained by two theories, *viz.,* Surprisal Theory (Levy, 2008) and DLT (Gibson, 2000), which define per-word complexity measures related to the above features. Further, to examine the effects of prosody we look into duration as a feature to explain disfluencies. We formulate duration as the time taken to utter the whole word but also examine the effects of two more prosody features - elongation (word duration/average word duration) and duration normalised by syllables. The following subsections describe these theories and our predictors based on them.

2.1 Surprisal Theory

Building on Shannon's (1948) definition of information, it has been shown in recent work formalized as Surprisal Theory (Hale, 2001; Levy, 2008) that the information content of a word is a measure of human sentence comprehension difficulty. The surprisal of a word is defined as the negative log of its conditional probability in a given context (either lexical or syntactic). We deploy lexical surprisal as measure of predicting disfluencies. We use the definition proposed by Hale (2001) which states that the lexical surprisal of the k^{th} word w_k in a sentence is

$$S_k = -logP(w_k \mid w_{k-1}, w_{k-2}).$$

Where $P(w_k \mid w_{k-1}, w_{k-2})$ refers to the conditional probability of k^{th} word in the sentence given the previous two words. We calculate lexical surprisal of each word in our corpus by training a simple trigram model over words on the Open American National Corpus (Ide and Suderman, 2004) using the SRILM toolkit (Stolcke, 2002). Going beyond simple lexical n-grams with direct counts syntactic surprisal is calculated using PFCGs where the probability of each word w_k is calculated by summing the probabilities of all trees T spanning words w_k to w_1 i.e.,

$$P(w_k, w_{k-1}...w_1) = \sum_T P(T, w_k, w_{k-1}...w_1)$$

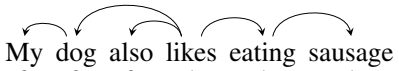
. Using this definition of probability, we define syntactic surprisal of the k^{th} word w_k as

$$S_k = -log\frac{\sum_T P(T, w_k, w_{k-1}...w_1)}{\sum_T P(T, w_{k-1}, w_{k-2}...w_1)}$$

We calculate syntactic surprisal in our corpus by training a PCFG parser over sections 2 to 21 of the Penn Treebank Corpus (Marcus et al., 1994) using the ModelBlocks software (an incremental implementation of the Berkeley parser).

2.2 Dependency Locality Theory: Integration and Storage Costs

Our second theory is the Dependency Locality Theory (henceforth DLT) proposed by Gibson (2000). The central notion of DLT revolves around two costs: integration cost (IC) and storage cost (SC), which have successfully accounted for the comprehension difficulty associated with many constructions (subject and object relative clauses for example). We depart from Gibson's definitions of these costs and compute DLT costs as follows: For a word to be integrated into the structure built so far, its integration cost, a backward-looking cost, would be the sum of the dependency lengths of all dependencies that include the word to be integrated and its previously encountered head/dependent word (grammatical link provided by dependency grammar). In contrast, the storage cost is a forward-looking cost and corresponds to the number of incomplete dependencies in our integrated structure thus far. To calculate these costs, the dependency relations for our corpus were extracted by removing disfluencies from the constituency-based parse trees and converting these trees into dependency graphs using the Stanford parser (De Marneffe et al., 2006). Though Gibson's original DLT formulation was based on constituency structures (with empty categories), we compute DLT costs from dependency parses for ease computation and assuming no empty categories. We illustrate the calculation of these DLT inspired costs in detail with the following example:



	My	dog	also	likes	eating	sausage
SC:	2	2	3	1	1	1
IC:	1	1	1	2	1	1

To calculate storage cost of a particular word, say, *also* we must calculate the number of incomplete dependencies that have a missing head in the structure so far, i.e. *My dog also*. In our case we see that there are 2 such incomplete dependencies, one from *likes* to *dog* and another from *likes* to

also. We add a minimum cost of 1 with the number of incomplete dependencies (2) to get a storage cost of 3 for *also*. For the case of *likes* the storage cost would be 1 (the minimum cost) because the incomplete dependency from *likes* to *eating* has a head (*likes*) that is already part of the structure. To calculate integration cost of a particular word, say, *likes* we must calculate the sum of dependency lengths from *likes* to the structure so far, i.e. *My dog also likes*. We see that there are two dependencies, the one from *likes* to *dog* having a length 1 and the other from *likes* to *also* having a length 0 (length is measured by number of intervening words). We then add the minimum cost 1 to the sum of dependency lengths to get a integration cost of 2 for *likes*.

3 Experiments and Results

In our study we use the version of the Switchboard corpus provided by the Switchboard in NXT project (Calhoun et al., 2010), which combines the annotations from the Penn Treebank3 corpus (Marcus et al., 1994) and MS-State transcripts (Deshmukh et al., 1998) along with adding new information like discourse and prosody. It consists of over 720,000 words and has a range of different sorts of linguistic information annotated on it including syntax, duration, discourse, and prosody. We focus on 3 classes of words taken from the Switchboard NXT corpus: reparandum, disfluent filler, and fluent word. For each of these classes we base our features for training a GLMM on the fluent words that immediately follow or precede the target (for reparandum-based disfluencies, these are taken as the words that immediately follow the repair and precede the reparandum). This was done for uniformity as disfluencies such as the disfluent filler *uh* do not posses the same linguistic features as fluent words. All the cases where the surrounding words have unclear POS tags or non-aligned duration have been excluded from this dataset. This results in a total of 14520 cases of reparandums, 12050 cases of disfluent fillers and 558361 cases of a fluent word. Further, to have balanced classes we randomly sample an equal number of fluent words for both types of disfluency, resulting in the following datasets for the two binary classification tasks: 29040 instances for *reparandum vs fluent* and 24100 instances for *filler vs fluent*.

In order to test whether our predictors are significant predictors of disfluencies (*i.e.,* fillers or reparandums), for the main results reported in this paper, we used the following GLMM implemented using the 'lme4' package in R:[1]

$$disfluency \sim lexsurp + synsurp + IC + SC + duration + (1|word)$$

Both the dependent variables above are binary choice (1-disfluency; 0-fluent). GLMMs can be thought of as a generalization of logistic regression models which allow for random factors as well as fixed factors. A random factor would mean that our model contains a separate intercept term for each category of that factor hence representing the features at a more individual level for these random factors. We set up our main GLMM with raw words as a random factor to control for the lexical variation in our model. The fixed factors are lexical surprisal, syntactic surprisal, DLT storage and integration costs, and word duration for all words (refer to Section 2 for actual computations). In the remaining subsections of this section, we describe our experiments with models encoding different random factors before finalizing the GLMM with raw words as the random effect term.

3.1 Random effects

In addition to a model with words as the random effect term, we examined the performance of GLMMs containing the fixed effects described in the previous section and the following random factors: fine-grained POS tags and coarse POS (collapsing nouns, adjectives, adverbs, and verbs to content words and the rest to function words). These choices are also supported by prior work in the disfluency detection where pattern matching and similarity measures of POS tags and words are shown to be effective feature types in detecting reparandums (Honnibal and Johnson, 2014; Zayats et al., 2016; Wang et al., 2017). Using a 5-fold cross-validation split in data we then trained our GLMMs and examined their disfluency classification accuracy on the entire dataset (Table 2). While the GLMM with raw words as random effects resulted in the best accuracy, models with coarse and fine POS random effects showed a considerable increase from the baseline setup of no random effects. A plausible explanation for their performance is that function words tend to occur disproportionately in disfluent contexts (Bell et al., 2003; Tree and Clark, 1997; Shriberg, 1995, refer Section 3.5

[1] We adopted the R GLM format for presenting the model: the dependent variable occurs to the left of '$\sim$' and independent variables occur to the right; 1| random factor

Features	Model with all features				Incremental models	
	Fillers		Reparandum		Filler	Reparandum
	Coef	Std Error	Coef	Std Error	Accuracy	Accuracy
Intercept	-1.34*	0.538	-0.73	0.376	66.80%	66.29%
1. Preceding Lexical Surprisal	-0.55***	0.022	-0.26***	0.017	66.81%***	66.22%
2. Following Lexical Surprisal	**0.74*** **	**0.023**	**0.32*** **	**0.017**	**71.84%*** **	**66.96%*** **
3. Preceding Syntactic Surprisal	-0.2***	0.025	0.09***	0.018	71.87%	67.02%**
4. Following Syntactic Surprisal	0.26***	0.023	-0.03	0.018	71.91%***	67.00%
5. Preceding Integration Cost	-0.09***	0.018	-0.06***	0.013	71.88%	67.01%**
6. Following Integration Cost	0.11***	0.022	0.23***	0.017	71.88%	66.96%***
7. Preceding Storage Cost	**-0.32*** **	**0.024**	**-0.47*** **	**0.018**	**72.26%*** **	**67.53%*** **
8. Following Storage Cost	**0.26*** **	**0.023**	**0.38*** **	**0.015**	**72.63%*** **	**69.08%*** **
9. Preceding Duration	**1.38*** **	**0.023**	**0.45*** **	**0.017**	**81.23%*** **	**71.12%*** **
10. Following Duration	0.07***	0.022	0.03	0.017	81.21%***	71.11%***

Table 1: GLMM regression (containing all features) and prediction results (when features are added incrementally with McNemar's significance over model in previous row); * p-value < 0.05, ** p-value < 0.01 and *** p-value < 0.001.

Random effect	Fillers	Reparandums
None	75.41%	65.09%
Coarse POS	76.99%	65.23%
Fine POS	79.01%	67.70%
Raw word	81.21%	71.11%

Table 2: Accuracies for binary classification (fluency vs. disfluency) models with different random effects.

for more details). So a model with coarse POS tags as random effects captures this association and predicts disfluencies to be more likely in contexts involving function words. Further, in a model with fine POS tags as a random effect, an analysis of cases where the magnitude of the random intercept is high revealed the following trends: pronouns follow fillers, and coordinating conjunctions precede disfluencies over fluent words (refer figure 1 in appendix A for details). Fine POS tags therefore seem to be capturing some extra information which helps with the classification task.

Finally, we note that the model with raw words as random effect displays the best accuracy and improves the accuracy over the baseline model of no random effects (5.8% increase for fillers and 6.02% for reparandums). To deal with unknown words, 165 words with frequencies less than 30 were coded as a separate category. Further, we observe no overfitting in our model with raw words as the random effect, as the training accuracies are 71.77% for reparandums and 81.91% for fillers which are very close to the test accuracies mentioned in Table 2. While looking at the distribution of the random intercept terms, we observed a high magnitude for intercepts for the words *know* or *mean*. Since disfluencies are known to signal discourse cues for listeners (Arnold et al., 2000, 2003), we explain these effects as a consequence of discourse markers like *you know* and *I mean*. In fact, we also find that the raw word intercepts focus on information about other discourse markers such as *so*, *because*, *then*, and *but*. We also observe that less frequent words (frequency < 30) have on average longer durations (412ms, compared to the overall mean duration at 274 ms) and longer word lengths (5.89 characters, with the overall mean value at 4.03 characters). Further, the random intercept values suggest that less frequent words are unlikely to precede disfluencies. In the remaining subsections of this section, we describe the impact of the various fixed effects in the same model on the two disfluency classification tasks.

3.2 Lexical Surprisal

In this section, we examine the results from the GLMM by interpreting its regression coefficients as well as its performance as a classifier in terms of prediction accuracy using 5-fold cross validation. To set up a baseline performance for the two binary classification tasks, we trained a classifier without any features but with raw words as random effects. This baseline accuracy comes out to be 66.80% for fillers and 66.29% for reparandums. To this baseline, we then start adding predictors based on surprisal, DLT, and duration incrementally until we get our final GLMM having all the features (81.21% accuracy for fillers and 71.11% for reparandums). Table 1 reports the regression coefficients (and their significance) of the final GLMM on the full feature set along with the incremental accuracies and the McNemar's significance (McNemar, 1947) of adding individual features one at a time. As evinced from the table, all lexical surprisal features turn out to be significant (regression coefficients with $p < 0.001$). Further, the lexical surprisal of words

following disfluencies ranks among the four best predictors in our set of features, inducing a significant accuracy increase (McNemar's two-tailed significance $p < 0.001$) of 5.03% for fillers and 0.74% for reparandums. The coefficients from Table 1 indicate that the words that follow both kinds of disfluencies (this would be the word following the repair in the case of reparandums) show a high lexical surprisal, suggesting that disfluencies occur in the presence of retrieval-based production difficulties. Previous studies have similarly shown that disfluencies occur in the presence of production difficulties due to new information (Arnold et al., 2000; Barr, 2001; Arnold et al., 2003; Heller et al., 2015). Examples from the corpus illustrated such behaviour in disfluent sentences such as "*for the **uh** scud missiles*" or "*imagine **that's a - that's a** pillsbury plant?*" having high surprisal words like *scud* or *pillsbury* following the disfluency. We also note that adding the lexical surprisal of the word preceding the disfluency leads to a significant increase in accuracy for fillers (McNemar's two-tailed significance $p < 0.001$). Further, the negative coefficient with lexical surprisal is suggestive of lesser retrieval difficulties for words before disfluencies, a theme we take up in the discussion in Section 4.

3.3 Syntactic Surprisal

In terms of classification accuracy, we note that syntactic surprisal does not give much improvement over lexical surprisal, except for the syntactic surprisal of words following reparandums. A perusal of the regression coefficients indicates that as with lexical surprisal, the words that follow disfluent fillers show high surprisal, suggesting that disfluencies occur in the presence of syntactic difficulties. We also see low syntactic surprisal for words preceding disfluent fillers, suggesting that syntactic difficulty is not heightened and is even lowered before fillers. Interestingly, we observed that reparandums act differently from fillers and report high syntactic surprisal for words that precede reparandums. This may be due to the fact that unlike fillers, reparandums consist of words in themselves and these words may be the ones that have low surprisal rather than the word preceding the reparandum. Similar effects have been observed in previous work on disfluencies by Dammalapati et al. (2019; 2020).

3.4 DLT: Integration and Storage Costs

We note that all the DLT costs have significant effects in the final GLMM (regression coefficients with p-value < 0.001 in Table 1). Further, we observed high integration and storage costs for words following disfluencies, indicative of upcoming difficulties in the case of disfluencies. On the other hand, words preceding disfluencies report lower DLT costs, indicating a lowering of difficulty (marked by low preceding DLT costs) before disfluencies to possibly help process the upcoming difficulty better. In terms of prediction accuracy, however, integration costs do not induce significant increases over the baseline model containing lexical and syntactic surprisal predictors. Prior work in sentence comprehension by Demberg and Keller (2008) has also shown integration cost to behave anomalously while predicting reading times and to act in the expected direction only for high values of dependency length.

We note that storage cost is one of the stronger predictors among our features and boosts the accuracy significantly (McNemar's two-tailed $p < 0.001$) over the baseline, by 0.75% for fillers and 1.55% for reparandums. Storage costs probably perform well because being forward-looking costs, they model upcoming difficulties effectively. A finer grained analysis of storage costs (refer figure 2 in appendix B for details) reveals that nouns and verbs have storage costs in all ranges (high, medium, and low), while conjunctions and prepositions have low storage costs. Pronouns tend to occupy medium storage costs, while adverbs and adjectives predominate high storage costs. Based on the idea of modeling difficulties some of these observed patterns have a bearing on the results as conjunctions often occur before disfluencies and pronouns often appear after disfluent fillers though noun, verb, and adverb distributions are not distinctive for disfluent contexts in particular.

3.5 Duration

Table 1 shows that the best predictor of disfluencies is the duration of words preceding disfluencies (substantial prediction accuracy increase of 8.66% for fillers and 2.16% for reparandums). Apart from the duration of words following reparandums, all the other duration features turn out to be significant ($p < 0.01$). We also note that speakers take a

Condition	Elongation		Word length		Function words	
	P	F	P	F	P	F
Fluent	0.87	0.91	3.79	3.98	63%	60%
Filler	1.43	1.00	3.91	3.97	66%	66%
Reparandum	0.98	0.93	4.32	4.45	51%	47%

Table 3: Mean elongation (ratio), word length (chars) and proportion of function words for preceding (P) and following (F) words in different conditions.

longer time on words following and preceding disfluencies, in concert with previous findings by Bell et al. (2003). Duration increases on words prior to disfluencies can be explained as a way to help in planning for future difficulties.

Similar to duration, we observe from Table 3 that words surrounding disfluencies also have higher word lengths. Further, the words of the reparandum/repair have the shortest word length (average of 2.94 characters), which is probably because they consist of the highest proportion of function words. To better understand the association of duration with disfluencies, we also looked at a predictor we call *elongation* (defined as word duration divided by average duration of the same word) which measures how much longer a particular word takes to pronounce than usual, *i.e.,* whether it is being stretched out by the speaker. We see that there is indeed an effect of words being stretched out in the context of disfluencies, which can be explained as taking time to plan for difficulties. We also note that this elongation effect is especially high in the case of words before fillers as they take 1.43 times longer than average, possibly explaining why preceding duration is such a strong indicator of disfluent fillers. The absence of such a prominent elongation effect for the surrounding words of reparandums is perhaps explained by the fact that the words in the reparandum/repair, i.e., the disfluency itself, have an elongation of 1.2. Further, though our study primarily measures duration as the time taken to utter the full word, we have also looked at normalised (per syllable) duration and elongation and observe very similar results to those with raw duration.

Table 3 also depicts that compared to fluent words, the proportion of function words relative to content words is higher before and after fillers. This supports previous findings that function words occur disproportionately often in disfluent contexts (Bell et al., 2003; Tree and Clark, 1997; Shriberg, 1995). Contrary to fillers, we observe that reparandums have a higher proportion of surrounding content words (i.e., fewer function words) compared to fluent words. We hypothesize that this is because the reparandum and repair themselves have a low proportion of content words: 74% of the words therein are function words. Previous work has also shown that content and function words make use of distinct modes of access in speech production (Garrett, 1975, 1980; Lapointe and Dell, 1989; Bell et al., 2009), making it important to look at their effects in the case of disfluencies. We discuss the implications of this proposal for understanding speech disfluencies in the next section.

4 Discussion

Our results indicate that in comparison to fluent words, words preceding disfluencies tend to have lower lexical surprisal and lower syntactic complexity (low syntactic surprisal and DLT costs). These disfluencies also tend to occur before upcoming difficulties *i.e.,* high lexical surprisal words with high syntactic complexity (high DLT costs and high syntactic surprisal except in the case of reparandums). Though words preceding reparandums do not show a lowering in syntactic surprisal, this could be attributed to the fact that reparandums themselves consist of words, which may be the ones that hold low syntactic surprisal costs, rather than the word preceding the reparandum. Thus reparandums behave almost similarly to disfluent fillers with differences possibly arising due to effects being present in the word choice of the reparandum, *i.e.,* in the disfluency itself rather than surrounding it. Among all our features, preceding word duration is the strongest predictor for disfluencies with an increase in accuracy of 8.66% for fillers and 2.16% for reparandums. We observe that the duration of words is longer when surrounding disfluencies and furthermore, the words near disfluencies (especially fillers) tend to get elongated, i.e. spoken longer than their mean duration.

Within the Levelt framework of speech production outlined at the outset, it follows that the rate of construction of phonologically encoded strings is limited by factors like syntactic complexity and lexical activations of the word while the rate of articulation would be limited by the complexity and number of syllables. Hence, when construction of these words are slow, speakers would try to coordinate these mechanisms of planning and articulation by modifying the duration of the words in order to maintain the flow of speech (Bell et al.,

2009). In the context of our results, we then propose that the duration of words preceding disfluencies is lengthened as a means to aid the planning of the following word which due to low activation levels (high surprisal values) and syntactic difficulties (high DLT costs) are slower to construct. We further make note that despite the lengthening of words before disfluencies to coordinate with the difficulties present ahead there is still a break down in the flow of speech with the interjection of reparandums and disfluent filler words resulting possibly because this short-term coordination of lengthening words is insufficient to accommodate for the construction difficulties ahead.

Production ease is often attributed to ease of retrieval of words from memory (Bock and Warren, 1985). Since more accessible words (more salience, more predictability) are known to be easier to retrieve and surprisal quantifies contextual predictability, higher surprisal values are indicative of difficulties in the retrievability due to the poor accessibility of words. These high surprisal difficulties work as a strong predictor especially for disfluent fillers as lexical surprisal alone boosts classifier accuracy by 5.04%. Since duration is a good predictor of disfluencies, this performance could also be explained by the correlation of 0.49 lexical surprisal maintains with word duration (maximum among all features). Further, we explain the lowering in surprisal before disfluencies as speakers choosing words which are easier to produce (lower surprisal implies a higher ease of retrieval) in order to plan for the upcoming production difficulties (marked by high surprisal) better.

In the context of our DLT-based measures, while we do not observe significant improvements in accuracy with integration cost we note that storage costs competes as one of the stronger predictors especially in the case of reparandums where it boosts classifier accuracy by 2.12% (stronger than lexical surprisal). We explain this performance by how storage costs is a forward looking cost and models the upcoming difficulties better because of this. Though we see similar behaviours with DLT costs and surprisal in modeling difficulty, there are differences between the two features both theoretically by how DLT unlike surprisal is not probabilistic and empirically by the poor correlation between surprisal and DLT-based features (maximum correlation of 0.151). Work by Demberg and Keller (2008) has also shown that DLT integration cost and surprisal are uncorrelated and complementary in nature. They attribute this effect to the fact that integration cost is a backward looing measure (prior material in memory is integrated to current), while surprisal is a forward looking cost as mentioned before. Hence, it is well suited to incorporate both these aspects of processing complexity in order to form a more complete theory.

Finally, we see that function words tend to be predominant in the words surrounding disfluent fillers and in the words of the reparandum/repair. We account for this in the light of evidence from the speech production literature (Garrett, 1975, 1980; Lapointe and Dell, 1989; Bell et al., 2009) which suggests that function words have a privileged mode of access and is distinct from content words (which tend to be more sensitive to activation levels). The early and influential Garrett model (Garrett, 1975, 1980) proposed that syntactic templates are selected with forms of function words filled in, and content words are accessed at a later stage, filling lexical slots in the templates. In the Extended Garrett Model of Lapointe and Dell (Lapointe and Dell, 1989), function words belong to syntactic fragments and are accessed via a feature-lookup procedure. Content words are accessed via network activation, filling slots in syntactic phrase structures. Recent results by (Bell et al., 2009) also show support for the Extended Garrett Model of Lapointe and Dell where function words have a special mode or modes of access, and the access of content words is sensitive to their activation levels. Thus speakers might be choosing these simpler words which are easier to construct around disfluent fillers and in the reparandum/repair to better plan for the upcoming difficulties ahead of disfluencies. Speakers might be making use of this privileged mode of access in the case of disfluencies to aid the coordination between planning and articulation. Hence, we conclude that in the context of disfluencies speakers are trying to handle upcoming difficulties by choosing words of higher activation (often function words) that are easier to produce to precede disfluencies and lengthening the duration of these preceding words to coordinate and maintain the flow between planning and articulation. Speakers regardless are unable to maintain this consistent flow resulting in disfluent fillers and reparandums to appear as interjections in the speech.

Acknowledgements

The authors acknowledge extramural funding from the Cognitive Science Research Initiative Department of Science and Technology (DO: DST/CSRI/2018/263). We are also indebted to the anonymous reviewers of SCiL 2021, AMLaP 2020 and NAACL-SRW 2019 as well as Sidharth Ranjan for their invaluable comments and feedback.

References

Jennifer E Arnold, Maria Fagnano, and Michael K Tanenhaus. 2003. Disfluencies signal theee, um, new information. *Journal of psycholinguistic research*, 32(1):25–36.

Jennifer E Arnold, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1):28–55.

Dale J Barr. 2001. Trouble in mind: Paralinguistic indices of effort and uncertainty in communication. *Oralité et gestualité: Interactions et comportements multimodaux dans la communication*, pages 597–600.

Alan Bell, Jason M Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, 60(1):92–111.

Alan Bell, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory, and Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, 113(2):1001–1024.

J. Kathryn Bock and Richard K Warren. 1985. Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, 21:47–67.

Kathryn Bock and Willem JM Levelt. 1994. *Language production: Grammatical encoding*. Academic Press.

Sasha Calhoun, Jean Carletta, Jason M Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language resources and evaluation*, 44(4):387–419.

Herbert H Clark and Thomas Wasow. 1998. Repeating words in spontaneous speech. *Cognitive psychology*, 37(3):201–242.

Samvit Dammalapati, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2019. Expectation and locality effects in the prediction of disfluent fillers and repairs in english speech. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 103–109.

Samvit Dammalapati, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2020. Effects of Duration, Locality and Surprisal in Speech Disfluencies for English. In *Proceedings of the 26th Architectures and Mechanisms for Language Processing Conference (AMLaP)*, Potsdam, Germany. University of Potsdam.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. In *Lrec*, volume 6, pages 449–454.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Vera Demberg, Asad B. Sayeed, Philip J. Gorinski, and Nikolaos Engonopoulos. 2012. Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 356–367, Stroudsburg, PA, USA. Association for Computational Linguistics.

Neeraj Deshmukh, Aravind Ganapathiraju, Andi Gleeson, Jonathan Hamaker, and Joseph Picone. 1998. Resegmentation of switchboard. In *Fifth international conference on spoken language processing*.

Merrill Garrett. 1980. Levels of processing in sentence production. In *Language production Vol. 1: Speech and talk*, pages 177–220. Academic Press.

Merrill F Garrett. 1975. The analysis of sentence production. In *Psychology of learning and motivation*, volume 9, pages 133–177. Elsevier.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, pages 95–126.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Daphna Heller, Jennifer E Arnold, Natalie Klein, and Michael K Tanenhaus. 2015. Inferring difficulty: Flexibility in the real-time processing of disfluency. *Language and speech*, 58(2):190–203.

Matthew Honnibal and Mark Johnson. 2014. Joint incremental disfluency detection and dependency parsing. *Transactions of the Association for Computational Linguistics*, 2:131–142.

Nancy Ide and Keith Suderman. 2004. The american national corpus first release. In *LREC*.

Jeremy G Kahn, Matthew Lease, Eugene Charniak, Mark Johnson, and Mari Ostendorf. 2005. Effective use of prosody in parsing conversational speech. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 233–240. Association for Computational Linguistics.

Steven G Lapointe and Gary S Dell. 1989. A synthesis of some recent work in sentence production. In *Linguistic structure in language processing*, pages 107–156. Springer.

Willem JM Levelt. 1992. Accessing words in speech production: Stages, processes and representations. *Cognition*.

Willem JM Levelt, Ardi Roelofs, and Antje S Meyer. 1999. A theory of lexical access in speech production. *Behavioral and brain sciences*, 22:1–38.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 114–119, Stroudsburg, PA, USA. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Benjamin Munson. 2007. Lexical access, lexical representation, and vowel production. *Laboratory phonology*, 9:201–228.

Janet Pierrehumbert et al. 2002. Word-specific phonetics. *Laboratory phonology*, 7.

José C Pinheiro and Douglas M Bates. 2000. Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, pages 3–56.

Gregory Scontras, William Badecker, Lisa Shank, Eunice Lim, and Evelina Fedorenko. 2015. Syntactic complexity effects in sentence production. *Cognitive Science*, 39(3):559–583.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Elizabeth Shriberg. 1995. Acoustic properties of disfluent repetitions. In *Proceedings of the international congress of phonetic sciences*, volume 4, pages 384–387.

Elizabeth Shriberg, Rebecca Bates, and Andreas Stolcke. 1997. A prosody only decision-tree model for disfluency detection. In *Fifth European Conference on Speech Communication and Technology*.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Trang Tran, Shubham Toshniwal, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Mari Ostendorf. 2017. Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information. *arXiv preprint arXiv:1704.07287*.

Jean E Fox Tree and Herbert H Clark. 1997. Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, 62(2):151–167.

Shaolei Wang, Wanxiang Che, Yue Zhang, Meishan Zhang, and Ting Liu. 2017. Transition-based disfluency detection using lstms. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2785–2794.

Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2016. Disfluency detection using a bidirectional lstm. *arXiv preprint arXiv:1604.03209*.

Victoria Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2014. Multi-domain disfluency and repair detection. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Simon Zwarts and Mark Johnson. 2011. The impact of language models and loss functions on repair disfluency detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 703–711. Association for Computational Linguistics.

A POS tags (random intercepts)

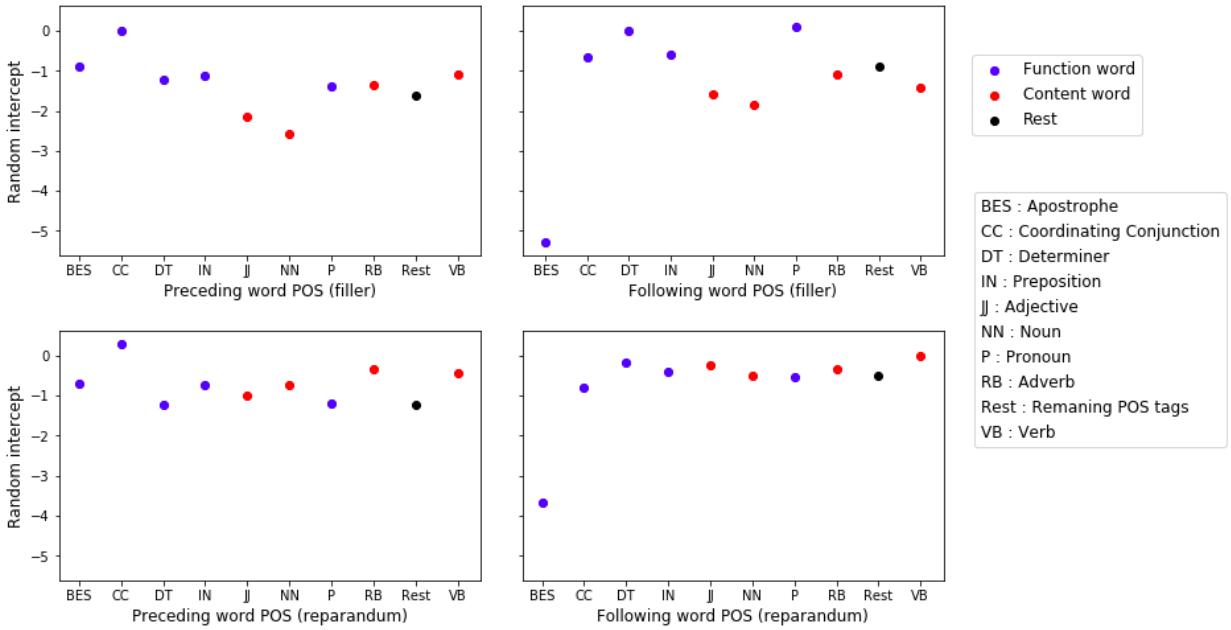


Figure 1: Scatter plot of random intercepts for different POS tags

B Storage Costs

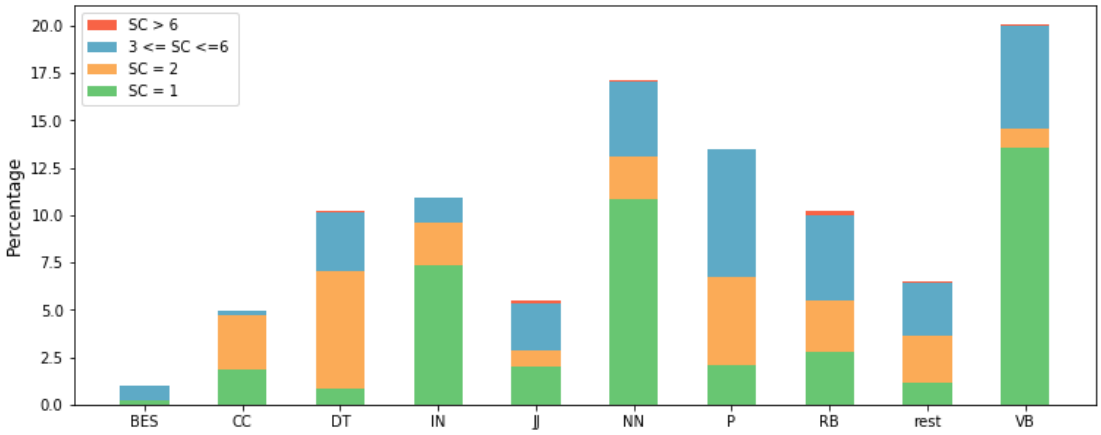


Figure 2: Percentage count of different POS tags split by Storage Costs