Effects of Duration, Locality and Surprisal in Speech Disfluencies for English

Samvit Dammalapati[1] Rajakrishnan Rajkumar[2] Sumeet Agarwal[1]
IIT Delhi[1], IISER Bhopal[2]
samvit1998@gmail.com, rajak@iiserb.ac.in, sumeet@iitd.ac.in

Speech disfluencies are typically classified into two categories: disfluent fillers and reparandums. Disfluent fillers are utterances like *uh*, *um* which do not have any meaning themselves but break fluency by interjecting and creating an interruption between words. Reparandums involve cases where speakers break fluency by making corrections or repetitions in their speech (examples can be found in the supplementary page). We focus on examining the role of two influential psycholinguistic theories, *viz.*, Surprisal Theory (Hale, 2001; Levy, 2008) and Dependency Locality Theory (DLT) (Gibson, 2000) as well as duration in accounting for disfluencies. Surprisal Theory defines an information-theoretic measure of comprehension difficulty *viz.*, surprisal. DLT proposes Storage and Integration costs to account for comprehension load at each word. Integration cost reflects the processing load associated with integrating a current word with previously encountered syntactic dependents. Storage cost encodes the load accrued while storing predictions about upcoming words using the number of incomplete syntactic dependencies in the integrated structure thus far.

In order to study these disfluencies, we use transcribed data from the Switchboard corpus (Godfrey et al., 1992), a corpus of fully spontaneous speech of American English. Using this corpus, our study focuses on examining 3 categories from it: reparandum, disfluent filler and fluent word. In order to test how the features *viz.*, surprisal costs, DLT costs and duration correlate with disfluencies we train a linear mixed effects models in a classification task to predict the disfluency category while providing the values of the features for the words preceding and following this category.

Our results as reported in Table 1 indicate that disfluencies tend to occur when the speakers have upcoming difficulties, as evinced from high DLT and surprisal costs at words following disfluencies. Speakers also seem to want to lower their cognitive load before disfluencies to help in planning, as suggested by low values of DLT and surprisal costs on the preceding word. Speakers also take a longer time on words following and preceding disfluencies which can be seen as way to help in planning for difficulties. The duration is measured as the time taken in utterance of the whole word. Even upon normalising the duration with syllable count, we note similar effects in the model. We also see that the means of normalised duration are higher for words that surround disfluencies (fluent - 0.20, filler - 0.27, reparandum - 0.23). Upon analysing the frequent words surrounding fillers and reparandum we note that around 55-60% are words similar to function words and do not have much content in and of themselves. The only exception being words following reparandum which showed only 40% of such type of words. Preliminary results also show a higher mean elongation (word duration/average word duration) for words that surround disfluencies (fluent - 0.89, filler - 1.21, reparandum - 0.96) suggesting speakers maybe elongating these words which are low in content.

We also see that the behavior of both categories of disfluencies, i.e., reparandums and fillers remain quite similar. Though the word preceding reparandums have a high syntactic surprisal which is against our expectations, this possibly arises due to a lessening of the cognitive load also happening in the word choice of the reparandum, i.e., in the disfluency itself. Similar effects in case of disfluencies have been observed in work by Dammalapati, Rajkumar, & Agarwal (2019). We see that despite similar behavior of DLT costs and surprisal, there are differences between the two features both theoretically by how DLT unlike surprisal is not probabilistic and emperically by poor correlation between surprisal and DLT (refer Table 2). This leads us to conclude that there is some independent value being added towards disfluency prediction by both surprisal and DLT measures.

Features	Fillers		Reparandum	
	Coef	**z-value**	**Coef**	**z-value**
Intercept	-1.24	-1.82	-0.95	-1.52
1. Preceding Lexical Surprisal	-0.43	-20.32***	-0.18	-11.04***
2. Following Lexical Surprisal	0.64	28.07***	0.17	10.43***
3. Preceding Syntactic Surprisal	-0.30	-12.59***	**0.02**	1.36
4. Following Syntactic Surprisal	0.34	15.56***	0.07	3.614***
5. Preceding Integration Cost	-0.05	-3.09**	-0.03	-2.03*
6. Following Integration Cost	0.04	1.68	0.23	13.72***
7. Preceding Storage Cost	-0.34	-14.33***	-0.51	-28.03***
8. Following Storage Cost	0.23	10.02***	0.33	21.86***
9. Preceding Duration	**1.34**	58.35***	**0.45**	25.97***
10. Following Duration	**0.12**	5.66***	**0.08**	4.40***

Table 1: Mixed effects model trained on all the mentioned preceding and following word features. * denotes p-value < 0.05, ** denotes p-value < 0.01 and *** denotes p-value < 0.001.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1.	1									
2.	-0.009	1								
3.	0.259	-0.007	1							
4.	-0.075	0.304	0.567	1						
5.	0.151	-0.014	0.025	-0.027	1					
6.	-0.007	0.147	-0.024	0.038	-0.04	1				
7.	-0.043	0.009	-0.066	-0.019	-0.261	0.275	1			
8.	-0.038	-0.052	0.008	-0.039	-0.065	-0.267	0.236	1		
9.	0.495	0.009	0.195	-0.029	0.16	-0.004	-0.13	0.002	1	
10.	-0.023	0.496	-0.002	0.209	0	0.173	-0.015	-0.179	0.053	1

Table 2: Correlation matrix for the features. Features are ranked in the same order as Table 1.

References

Dammalapati, S., Rajkumar, R., & Agarwal, S. (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*, (pp. 103–109).

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, (pp. 95–126).

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

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

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

For the computational modelling aspect of this study, we use the corpus provided by the switchboard NXT project and focus on looking at 3 categories from the Switchboard corpus: reparandum, disfluent filler and fluent word. Disfluent fillers are utterances like *uh*, *um* which do not have any meaning themselves but break fluency by interjecting and creating an interruption between words. For example, suppose a speaker says *thinking about the uh day when I*. Here, there is a break of fluency between the words *the* and *day* due to the interjection of the filler *uh*. Reparandums involve cases where speakers break fluency by making corrections or repetitions 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-* by restarting with the intended (corrected) speech *to the left*. We call the words to be corrected as the reparandum (*to the righ-*) and the correction the speaker follows with as the repair (*to the left*).

We base our features for modelling from the fluent words that immediately follow or precede these disfluencies (for reparandums these are taken as the words that immediately follow repair and precede the reparandum), this was done out of uniformity as disfluencies such as a disfluent filler *uh* do not posses the same linguistic features as fluent words. This results in a total of 14923 cases of reparandum, 12183 cases of disfluent filler and 558361 cases of a fluent word. For all these cases in the dataset we take the words preceding and following them and calculate features pertaining to POS tags, lexical surprisal, syntactic surprisal, DLT storage and integration costs, and word duration. Using linear mixed-effects modeling, we examine how these features behave in predicting disfluencies.

Linear mixed effects can be thought of as a generalization of linear regression which allows 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 linear mixed effects models with two random factors – speakers and POS tags. By doing this we control for the speaker variation in our model and also because disfluency detection has found pattern matching and similarity measures of POS tags to be effective feature types in detecting reparandums.

We train individual linear-mixed effects models to classify different disfluency types, i.e. we train two mixed-effects classifiers where one classifies disfluent fillers versus fluent words, and the other classifies reparandums versus fluent words. To prevent a data imbalance in our training data for the classifier we do a random sampling to ensure equal class sizes. For example, when we classify between reparandums and fluent words we randomly sample an equal proportion from fluent words i.e., 14923 cases of reparandum and fluent words in our training data. After doing this, we set up a baseline model with the random intercepts as speakers and POS tags. We measure the model fit using Akaike Information Criterion (AIC) and find that the model fit gets better (lower AIC) on adding these random effects. The baseline model setup had an AIC of 29735 for fillers and 36915 for reparandums. Further, successively adding features like surprisal, DLT costs and duration onto the mixed-effects model gives it a better fit. The resulting mixed-effects model with all the features added into it is described in Table 1. The AIC scores for this final mixed-effects model came out to be 22250 for fillers and 34643 for reparandum.