Expectation and Locality Effects in the Prediction of Disfluent Fillers and Repairs in English 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 three influential theories of language processing, *viz.*, Surprisal Theory, Uniform Information Density (UID) hypothesis and Dependency Locality Theory (DLT), in predicting disfluencies in speech production. To this end, we incorporate features based on lexical surprisal, word duration and DLT integration and storage costs into logistic regression classifiers aimed to predict disfluencies in the Switchboard corpus of English conversational speech. We find that disfluencies occur in the face of upcoming difficulties and speakers tend to handle this by lessening cognitive load before disfluencies occur. Further, we see that reparandums behave differently from disfluent fillers possibly due to the lessening of the cognitive load also happening in the word choice of the reparandum, i.e., in the disfluency itself. While the UID hypothesis does not seem to play a significant role in disfluency prediction, lexical surprisal and DLT costs do give promising results in explaining language production. Further, we also find that as a means to lessen cognitive load for upcoming difficulties speakers take more time on words preceding disfluencies, making duration a key element in understanding disfluencies.

1 Introduction

In contrast to written text which can be rewritten or edited, speech happens spontaneously making it more prone to mistakes. Speakers tend not to speak fluently and take pauses or even repeat words. Such errors where speakers interrupt their flow of speech are known as disfluencies. One of the primary reasons for speech disfluencies is difficulties in language production (Tree and Clark, 1997; Clark and Wasow, 1998). In this study, we aim to understand the role of disfluencies and classify disfluencies into two categories namely, disfluent fillers and reparandums. Disfluent fillers are utterances like *uh*, *um* which 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 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*).

In order to study disfluencies, we use transcribed data from the Switchboard corpus (Godfrey et al., 1992), a corpus of fully spontaneous speech of American English. We focus on testing the role of three influential linguistic theories, *viz.*, Surprisal Theory (Levy, 2008; Hale, 2001), Uniform Information Density (UID) hypothesis (Jaeger and Levy, 2007) and Dependency Locality Theory (Gibson, 2000) in accounting for disfluencies. Surprisal Theory defines an information-theoretic measure of comprehension difficulty *viz.*, surprisal. Recently, Demberg et al. (2012) showed that syntactic surprisal is a significant predictor of word duration in spontaneous speech even amidst the presence of competing controls like lexical frequency. Thus surprisal can be used to model language production as well, with words with high surprisal associated with speech disfluencies *i.e.*, fillers and repairs. The UID hypothesis predicts that in language production, speakers prefer to minimize variation of information density (mathematically same as surprisal) across the speech signal. Thus based on the UID hypothesis, it is plausible to assume that disfluencies are associated with higher informa-

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 103–109
Minneapolis, Minnesota, June 3 - 5, 2019.

tion density variation. Finally, DLT posits integration and storage costs as measures of comprehension difficulty. Scontras et al. (2015) showed that for English relative clause production, locality results in greater speech disfluencies and starting time for object relatives compared to subject relatives. Thus, we conceive higher values of integration and storage costs leading to disfluencies in language production.

We predict disfluencies in the Switchboard corpus using a one-vs-all logistic regression classifier containing features based on lexical surprisal, UID, DLT-inspired costs and duration. Further, by looking into the classifier's regression weights and accuracies, we get an insight behind how these theories affect disfluencies in speech. Our results do not uncover evidence to indicate UID hypothesis plays a significant role in disfluency prediction; however, lexical surprisal and DLT costs do give promising results in explaining language production. The latter two theories indicate that disfluencies tend to be followed with upcoming difficulties and speakers lower cognitive load on words preceding these disfluencies to ease this difficulty. Apart from these three theories, we look into how duration behaves in disfluent contexts and find that speakers take more time in words preceding disfluencies which we explain as a means to lower cognitive load for upcoming difficulties by buying more processing time. Further, we see that reparandums do not occur prior to words with lower surprisal like in the case of disfluent fillers. This effect may be due to the lessening of the cognitive load also happening in the word choice of the reparandum, i.e., in the disfluency itself.

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 the same or 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 collected features from lexical language models, dependency grammar and prosody, we are motivated to test influential linguistic theories in these domains and study whether disfluencies can be explained by these theories, *viz.,* Surprisal Theory (Levy, 2008), the UID hypothesis (Jaeger and Levy, 2007) and DLT (Gibson, 2000). Further, to examine the effects of prosody we look into duration as a feature to explain disfluencies.

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). The second theory we examine, the Uniform Information Density (henceforth UID) hypothesis also relates to information density and states that language production exhibits a preference for distributing information uniformly across a linguistic signal. Our third theory is the Dependency Locality Theory (henceforth DLT) proposed by Gibson (2000). This theory defines processing costs that have successfully accounted for the comprehension difficulty associated with many constructions (subject and object relative clauses for example).

3 Experiments and Results

Our study focuses on 3 classes of words in the Switchboard corpus: reparandum, disfluent filler and fluent word. We use the corpus provided by the switchboard NXT project (Calhoun et al., 2010) and base our features for machine learning from the fluent words that immediately follow or precede these disfluencies (for reparandum based disfluencies, 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

Features	Accuracy	Fluent	Filler	Repar.
Baseline	37.18%			
Preceding surprisal**	38.12%	0.014	-0.0664	0.0525
Following surprisal**	38.78%	-0.1204	0.0915	0.0389
Both surprisals**	40.02%			

Table 1: Accuracy and weights for lexical surprisal. Here * denotes p-value < 0.05 and ** denotes p-value < 0.01 for McNemar's test relative to the baseline.

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 14923 cases of reparandum, 12183 cases of disfluent filler and 558361 cases of a fluent word. For uniformity in classes we randomly sample 12183 cases from each class. We setup a one-vs-all logistic regression classifier to classify between our 3 categories. To set up a baseline performance for this multi-classification, we train the classifier on features pertaining to the speaker and listener particularly the gender, age and rate of speech. The change in accuracy relative to the baseline on adding features, *viz.*, lexical surprisal, UID, DLT costs and duration, tells us whether these theories explain the presence of disfluent contexts. Further, using the regression weights we look at whether the correlations are indeed as the theory expects. The next three subsections will describe these results in depth.

3.1 Lexical Surprisal

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). The lexical surprisal feature of the surrounding words turns out to be significant with a p-value < 0.05 and we can note from Table 1 that these classifiers with surprisal give a significant improvement from baseline (McNemar's test). Further, using surprisal of the word following the disfluency gives a 1.6% boost in accuracy and the regression coefficients from Table 1 indicate that the words that follow disfluencies (this would be the word following the repair in the case of reparandums) show a higher surprisal, suggesting that disfluencies occur in the presence of an upcoming difficulty. Previous studies have shown similar results 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 a behaviour where disfluent sentences such as "*for the **uh** scud missiles*" or "*imagine **thats a - thats a** pillsbury plant?*" have high surprisal difficulties like scud or pillsbury following the disfluency.

We also note that lexical surprisal of the word preceding the disfluency leads to an accuracy increase of 0.94%. There is a low surprisal of the preceding words in the case of disfluent fillers, as seen from Table 1, suggestive that speakers use easier words (lesser cognitive load due to low surprisal) to handle the production problems better. However, the other type of disfluency, reparandum shows a higher preceding surprisal which might be attributed to the fact that unlike fillers, reparandums consist of words in themselves and these words may be the ones that hold the low surprisal rather than the preceding word to the reparandum.

3.2 Uniform Information Density (UID)

In order to quantify the uniformity in information density spread, we use two types of UID measures proposed in previous works by Collins (2014) and Jain et al. (2018). The two measures are as follows:

- $UIDglob = \frac{-1}{N}\sum_{i=1}^{N}(id_i - \mu)^2$
- $UIDglobNorm = \frac{-1}{N}\sum_{i=1}^{N}(\frac{id_i}{\mu} - 1)^2$

Here, N is the number of words in the sentence, id_i refers to the information density, i.e., lexical surprisal of i^{th} word and μ is the average information density of the sentence. We note from the equation above that the uniformity measure, UIDglob is defined as the negative of variance of lexical surprisal. Further, our second measure UIDglobNorm is nothing but the normalized version of the first measure UIDglob.

We calculate the UID measures for the surrounding words, i.e., the immediate preceding and following words to the class, and the UID measures for that sentence. From the accuracies that are listed in Table 2, we can see that the best accuracy is an increase of 0.42% from UIDglob of the

Features	Accuracy
Baseline	37.18%
UIDglob surrounding	37.22%
UIDglob sentence*	37.60%
UIDglob both*	37.49%
UIDglobNorm surrounding	37.31%
UIDglobNorm sentence*	36.94%
UIDglobNorm both*	37.28%

Table 2: Accuracy for UID measures.

Features	Accuracy	Fluent	Filler	Repar.
Baseline	37.18%			
Preceding IC**	37.65%	0.0042	-0.0069	0.0027
Following IC**	39.06%	-0.0334	-0.0375	0.0709
Preceding SC**	39.48%	0.2895	-0.0575	-0.232
Following SC**	37.56%	-0.1731	0.0098	0.1633
Both ICs**	39.18%			
Both SCs**	41.61%			
Both ICs and SCs**	48.57%			

Table 3: Accuracy and weights for DLT costs.

sentence. However, this UIDglob measure for sentence when normalized (UIDglobNorm) shows a net decrease in accuracy. Though these UID measures on the sentence are significant features with a p-value < 0.05, the UID measures on the surrounding words are not. Further, these improvements in accuracy upon using sentence level UID features are significant (McNemar's test) only with p-value 0.05 but not with 0.01. We see that the UID measure for the surrounding words causes an increase of 0.13% which is far less compared to the increase in preceding (0.94%) and following (1.6%) surprisal noted earlier. The UID hypothesis we've examined has hence failed to bring about any notable improvements in our model in comparison to competing explanations like surprisal theory. It may also be possible that the UID hypothesis is limited only to syntactic reduction in English (Jaeger and Levy, 2007). Previous work by Jain et al. (2018) in Hindi word order choices has shown that the UID measures does not outperform lexical surprisal.

3.3 DLT: Integration and Storage Costs

The central notion of DLT revolves around two costs: integration cost (IC) and storage cost (SC) as proposed by Gibson (2000). We 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.). Our theory of DLT makes use of these dependency parses and in this way is inspired from the original DLT that makes use of constituency parsing. We see from the McNemar's test (Table 3) that the increase in accuracy for DLT costs are all significant improvements w.r.t baseline. The integration costs and storage costs, all significant features with p-value < 0.05, give an increase of 2% and 4.43% individually (from Table 3). Further, combining all these four features gives a far larger increase of 11.39% from the baseline. This can indeed be explained as the two costs serve complementary functions (can be seen from their negative correlation of -0.26) as forward and backing looking costs, and a combination would possess greater information on the whole. We see that DLT seems to perform well for our disfluency prediction task and so we will proceed to examining the correlations of the DLT costs in disfluent contexts.

From Table 3 we see that following integration cost is expectantly high for reparandum based disfluencies but contrarily is lower in the context of disfluent fillers which goes against the expectation of an upcoming difficulty in the case of disfluencies. Recent work by Demberg and Keller (2008) has shown integration cost to behave anomalous and act in the expected direction only for high values of dependency length. With preceding word's integration cost getting lowered in the context of disfluent fillers and higher in the context of reparandums, we can see that apart from the anomalous result for following integration cost in fillers, integration cost functions gets explained similar to lexical surprisal. We also see that disfluencies, i.e., both disfluent fillers and reparandums have a lower preceding storage cost and higher following storage cost. This makes sense as a lower preceding storage cost lowers the cognitive load and helps process the upcoming difficulty better (indicated by high following storage cost).

3.4 Duration

Looking into how duration affects disfluencies, from Table 4, we can note that the duration of the preceding word gives a huge accuracy increase of

Features	Accuracy	Fluent	Filler	Repar.
Baseline	37.18%			
Preceding duration**	46.82%	-2.9315	2.7944	0.1371
Following duration**	37.71%	-0.4512	0.0046	0.4466
Both durations**	46.89%			

Table 4: Accuracy and weights for duration features.

9.64% from baseline. From McNemar's test it is indicated that the increase in accuracy from using duration features w.r.t baseline are all significant. The duration features are also significant features with p-value < 0.05 and are higher in the context of disfluencies as can be seen from the regression weights in Table 4. These results are in concert with Bell et al's (2003) study of duration in disfluent contexts. The higher duration of words preceding disfluencies suggests that speakers try to buy time in order to better process for the upcoming production difficulties that follow these disfluencies.

3.5 Correlations between features

We observe that the maximum correlation from the feature correlations is between duration and surprisal. This positive correlation of 0.49 between surprisal and duration can be expected as higher information density for a word would take the speaker a longer duration to process. Given the performance of duration in disfluency prediction, this correlation could also explain the significant performance of disfluency prediction with surprisal. Further, recent work by Demberg et al. (2012) has shown how syntactic surprisal is a significant predictor of word duration. This is suggestive of the fact that surprisal can possibly be used to model language production, despite being an information-theoretic measure of comprehension difficulty. For further correlation values between features refer to Table 1 in Appendix A.

4 Discussion

Our results indicate that disfluencies occur when speaker has upcoming difficulties, as evinced from high storage cost and lexical surprisal 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 duration, storage cost, surprisal and integration cost on the preceding word. Ease of production is often attributed to ease of retrieval of words from memory. More accessible words (more salience, more predictability) are known to be easier to retrieve compared words with low accessibility. Since surprisal quantifies contextual predictability, the low surprisal prior to fillers indicate the ease of retrievability of words prior to fillers. Though words preceding reparandums do not show a lowering in surprisal and integration cost, it could be attributed to the fact that reparandums in itself consists of words, which may be the ones that hold low surprisal or integration costs, rather than the word preceding to the reparandum. This difference in the context of fillers and reparandums indicates the presence of distinct memory operations in language production.

DLT-based costs hitherto used to explain language comprehension gave the best increase in accuracy to 48.57%, showing it has promise in explaining language production too. We did however note that the integration costs behaved in contrary directions, and so further detailed research is needed. The anomalous DLT effects need to be investigated more thoroughly in future work. In the comprehension literature Vasishth and Lewis (2006) proffer a unified explanation for both locality and anti-locality effects in Hindi verb-final constructions by resorting to either decay or interference (on account of similar intervening words) at a verbal head while integrating a previously encountered argument head. In a survey of dependency distance, Liu et al. (2017) state that long dependencies might not be difficult to process due to the presence of mitigating factors like frequency, contextual familiarity and positional salience.

Given that DLT costs bring about a large increase in accuracy, incorporating other syntax-based features like syntactic surprisal and UID (based on syntactic surprisal) might confer further insights on the role of syntax in disfluency prediction and language production. Despite seeing how DLT, duration and lexical surprisal behave individually, we have not as such compared these features against each other and studied if they account for explaining same parts of the data. We leave these steps for future work. Our modelling presupposes linear dependence between individual predictors. In future inquiries we plan to use other non-linear classifiers like decision trees and KNN models.

[0]We thank the anonymous reviewers, Micha Elsner and Sidharth Ranjan for insightful 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, 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.

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.

Michael Xavier Collins. 2014. Information density and dependency length as complementary cognitive models. *Journal of psycholinguistic research*, 43(5):651–681.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses.

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.

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*.

T Florian Jaeger and Roger P Levy. 2007. Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems*, pages 849–856.

Ayush Jain, Vishal Singh, Sidharth Ranjan, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2018. Uniform information density effects on syntactic choice in hindi. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 38–48.

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.

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

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171 – 193.

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, 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.

Shravan Vasishth and Richard L. Lewis. 2006. Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, 82(4):767–794.

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.