Modeling Bilingual Disfluencies with Large Language Models

Negin Raoof [*1] **Yating Wu** [*1] **Carlos G. Bonilla** [*1] **Junyi Jessy Li** [2] **Stephanie Grasso** [3] **Alexandros G. Dimakis** [1] **Zoi Gkalitsiou** [3]

Abstract

Speech disfluency metrics are commonly used for informing diagnosis and treatment of various communication disorders. However, bilingual speakers exhibit unique speech disfluency patterns, increasing the difficulty of speech and language disorder diagnosis in bilingual children and adults.

We propose and train models for predicting disfluencies in monolingual and bilingual speakers, using LLMs and a modern machine learning pipeline. We use a novel bilingual dataset with detailed annotated disfluencies and participant information. We find that disfluencies tend to happen at high surprisal words, validating surprisal theory for both monolinguals and bilinguals. We also find some interesting differences in the manifestation of disfluencies between bilingual and monolingual speakers.

1. Introduction

Speech-language pathologists use disfluency metrics for informing diagnosis and treatment of various communication disorders including stuttering, aphasia and dementia. Approximately twenty percent of the US population is bilingual and unfortunately, bilingual children and adults are largely evaluated using disfluency measures designed for monolingual English speakers. This places them at a higher risk for misdiagnosis of speech or language disorders (Byrd, 2018; De Lamo White & Jin, 2011; Fabiano-Smith & Hoffman, 2018; Grasso et al., 2023). Despite the well-documented cognitive and socio-cultural benefits associated with bilingualism, biases against bilinguals and misdiagnoses of communication disorders in this group leads to unique challenges.

Neurotypical monolingual speakers exhibit different speech disfluency patterns compared to bilinguals. This increases the difficulty of accurate diagnosis of speech and language disorders in bilingual children and adults. Speech-language pathologists commonly report challenges in evaluating children and adults who speak more than one language (e.g. (Boerma & Blom, 2017; Byrd et al., 2016; Grimm & Schulz, 2014; Hemsley et al., 2014)).

In this paper we propose and train a model for predicting disfluencies for bilingual speakers, using LLMs. Disfluencies are believed to be caused by challenges in planning, or failing to coordinate speech planning and execution (Bell et al., 2009; Dammalapati et al., 2021). Our model builds on two influential theories of language processing: Surprisal theory (Levy, 2008; Hale, 2001) and Dependency Locality Theory (DLT) (Gibson, 2000). Recent work by (Dammalapati et al., 2021) developed a disfluency prediction model that used surprisal and DLT metrics, and studied which syntactic and information theoretic features are predictive of different types of disfluencies in monolingual English speakers. We follow a similar methodology but also study the differences of the learned disfluency prediction models for monolingual and bilingual speakers. Our contributions are as follows:

1. We create a novel bilingual dataset with annotated disfluencies and detailed participant information measures. Our dataset also includes measures of bilingualism per participant including age of acquisition of each language, years lived in the US, percentage of use of each language at the time of the study, and self-reported proficiency in each language in the areas of speaking, comprehension, reading and writing.

2. We validate surprisal theory using GPT-2 (Radford et al., 2019) as a model to estimate conditional (next token) probabilities.

3. We study the differences in the manifestation of disfluencies in bilingual vs. monolingual speakers by training two different models and comparing their coefficients. We find that the predictive value of lexical and phonetic features on disfluencies is different for

[*]Equal contribution [1]Department of Electrical and Computer Engineering, University of Texas at Austin, Texas, USA [2]Department of Linguistics, University of Texas at Austin, Texas, USA [3]Department of Speech, Language, and Hearing Sciences, University of Texas at Austin, Texas, USA. Correspondence to: Negin Raoof <neginmr@utexas.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

monolingual vs bilingual speakers.

2. Method

2.1. Participants

In this study we collected data from two different groups of participants: 25 Greek-English bilingual adult speakers (avg=27.42, SD=5 years; age range= 20-46 years) and 15 adult speakers who spoke only English (avg=28.9, SD=11.5 years; age range= 21-62 years).

2.2. Dataset

2.2.1. Participants

In this study we collected data from two different groups of participants. The bilingual group included 25 Greek-English bilingual adult speakers (avg=27.42, SD=5 years; age range= 20-46 years) and the monolingual group included 15 adult speakers who spoke only English (avg=28.9, SD=11.5 years; age range= 21-62 years).

The bilingual participants were all sequential bilinguals, who spoke Greek as their first language and learned English at or after 4 years of age (age range: 4-12 years old). They all moved to the US as adults and had lived in the US for at least a year at the time of the study. Based on their reports, nine participants reported using each language 40%-60% of the time during a week and the remaining 16 reported that they used English more than 60% of the time. Language proficiency was measured through participants' self-ratings of each language in the areas of speaking, comprehension, reading and writing, using a 1-7 point scale (1= not proficient at all, 7= completely/native-like proficiency). Participants reported high proficiency in both languages. Specifically, the mean rating scores for the participants' proficiency in Greek were: speaking= 6.92 (SD= .27), comprehension= 6.96 (SD= .20), reading= 6.96 (SD= .20), and writing= 6.85 (SD= .46). In English, the mean rating scores reported were: speaking= 5.08 (SD= 1.02), comprehension= 5.88 (SD= .82), reading= 6.12 (SD= .86), and writing= 5.35 (SD= 1.06).

2.2.2. Disfluency Data

This study utilized two datasets consisting of narrative speech samples produced in English from non-native Greek-English bilingual speakers (bilingual dataset) and native English monolingual speakers (monolingual dataset). The bilingual dataset included 16591 samples, each sample representing a produced word. The monolingual dataset included 10896 samples. Each participant produced two narratives: one picture description and one story-telling based on a wordless picture book. This textual dataset carefully identified disfluencies that were annotated by trained speech-language pathologists. The disfluencies were then classified into three categories: between-word disfluencies (BWD), within-word disfluencies (WWD), and repetitions (REP).

BWD included reparandums (e.g., phonological fragments and word or phrase revisions/repairs) and fillers (e.g., filled pauses, or interjections, hesitations etc.).

WWD included broken words and repeated word segments (e.g., sound or syllable repetitions).

REP included whole-word repetitions (once or many times) and short phrase repetitions (once or many times).

2.3. Prediction Task

In our study, we trained (logistic regression) classifier models to predict the occurrence of BWD, WWD and Repetitions (REP) in monolingual and bilingual speakers' transcribed speech. Our objective was to identify and compare the features that effectively predict disfluencies among these two participant groups. We examined the significance of model features to compare predictive attributes for each disfluency type in the two language groups.

We had one binary classifier model trained per each disfluency type (BWD, WWD, and REP) and separate models were trained for both monolingual and bilingual datasets.

2.4. Features

2.4.1. Language Model Features

Surprisal Surprisal theory (Levy, 2008; Hale, 2001) introduced the information theoretic concept of surprisal as a predictor of cognitive load and previous recent work showed that it indeed is predictive of disfluencies (Dammalapati et al., 2021).

The surprisal of a word is therefore defined as the negative logarithm of its conditional probability in a given context of previous words. While the context can be lexical or syntactic, we use lexical surprisal, as defined by (Hale, 2001), which states that the lexical surprisal of the k^{th} word w_k in a sentence is

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

2.4.2. Syntactic Features

Integration and storage costs The two important features to capture the syntactic complexity of a sentence are explained by the Dependency Locality Theory (DLT) (Gibson, 2000). The theory evaluates complexity using two metrics: integration cost and storage cost. Integration cost (IC) is determined by summing the dependency lengths of all prior head/dependent words. Storage cost (SC) is the count of incomplete dependencies given the current segment. Note that this metric

depends on the whole utterance, and not only on context.

An example of the computation of DLT costs can be shown in Fig 1. To compute SC for a word w_i, first we create the set of all words which appear before w_i in the utterance. Then, we count the number of arrows entering this set and add 1 to this number (that ensures that the minimum cost is 1). To compute IC for a word w_i, first we create the set of all words which appeared before, including w_i, in the utterance. Define the length of a directed edge to be the number of words between the head and tail of that edge. The IC is the sum of the lengths of the edges in the set connected to w_i plus one. Note that the connections can be in either direction.

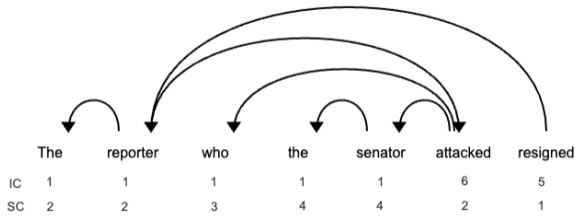

Figure 1. DLT costs computation. For example for the word $w_i =$ 'attacked' the set of previous words includes 'The reporter who the senator attacked'. To compute the IC we need to sum the lengths of the paths connected to 'attacked'. There are three such paths connecting 'attacked' with 'reporter', 'who' and 'senator' and their corresponding lengths are 3, 2, 0. Therefore, the IC for $w_i =$ 'attacked' is $3 + 2 + 0 + 1$. To compute the SC, we count the number of incomplete arrows entering the set. There is one incomplete arrow entering the set from 'resigned'. Therefore, the SC for $w_i =$ 'attacked' is $1 + 1$.

2.4.3. Lexical Features

We use a set of lexical features recognized to be associated with speech disfluencies.

Word Frequency Word frequency refers to how often a word is used and affects processing speed. It usually relies on counts of written corpora and the most common standardized measure is frequency per million of words ((Brysbaert et al., 2018)).

Neighbor Density Neighborhood density refers to the number of phonological or orthographic neighbors a word has by adding, deleting, or changing a sound in the word's phonetic representation.

Stop Word We used the spacy library to identify if a word was a stop word. Stop words are a closed class of words and include prepositions, pronouns, conjunctions, determiners and auxiliary verbs. They are considered less complex as they do not carry any meaning and are usually shorter in length.

Content Word We used the nltk library to do part-of-speech (POS) tagging and identify content words accordingly. Content words are an open word class and their main role is to convey meaning. They include nouns, verbs, adjectives and adverbs.

2.4.4. Phonetic Complexity Features

The following features were chosen based on their relevance in speech production.

IPC The Index of Phonetic Complexity is a phonetic complexity measure of a word. Higher IPC scores indicate higher complexity. According to (Jakielski, 1998), IPC is sum of the following eight components: consonants categorized by place, consonants grouped by manner, vowels classified by type, word shape, word length in syllables, singleton place diversity, contiguous consonants, and cluster variety.

Phonotactic Probability Phonotactic probability refers to how frequently a phonemic segment (or segment sequences) occur in a language ((Jusczyk et al., 1994)). In our study, we phonetically transcribed and translated all words into a computer-readable transcription (i.e., Klattese) and obtained their bigram phonotactic probabilities using the (Vitevitch & Luce, 2004) online calculator. For instance, when analyzing the English word "blick", we counted the frequency of pairs such as [bl], [lɪ] and so on.

2.5. Predictions and Evaluation Metrics

We pre-processed our dataset to compute all the aforementioned features. Then, we trained a logistic regression model with ablation of those features. We evaluated our model performance on a holdout set. Note that we split the training and test datasets according to participants.

We interpreted the model's coefficients as the relative importance of the feature. A larger coefficient was indicative of higher importance in predicting the dependent variable. We used McNemar's test to evaluate the significance of each predictor feature in our model.

3. Results

3.1. Surprisal Calculated by LLMs outperforms N-gram's probability

We computed lexical surprisal using probabilities from both GPT-2 and N-gram models, and compared the accuracy of

models in predicting disfluencies. Our findings revealed that when computing surprisal, leveraging GPT-2 yields more accurate results compared to the N-gram model (with N=3) for both monolingual and bilingual speakers. Results are presented in Table 1.

Dataset	Model	AUC
Monolinguals	GPT-2	0.723 ± 0.014
Monolinguals	N-gram Model	0.682 ± 0.016
Bilinguals	GPT-2	0.705 ± 0.010
Bilinguals	N-gram Model	0.676 ± 0.008

Table 1. Comparison of the performance of different models as source of probability for surprisal. We show the average AUC across test-set participants, also averaged across all types of disfluencies, plus corresponding standard error.

3.2. Between-Word Disfluencies (BWD)

According to (Dammalapati et al., 2021), words following disfluencies have high lexical surprisal and high syntactical difficulty, while words preceding disfluencies tend to have low lexical surprisal and low syntactic difficulty. We observed a similar trend in our monolingual and bilingual models, even though only surprisal following was significant. Lexical surprisal of words following disfluencies improved the monolingual model by 15.34 percent and the bilingual by 8.34 percent.

3.3. Within-Word Disfluencies (WWD)

Regarding WWD, we observed that lexical surprisal on disfluent words was the best predictor in the monolingual as well as the bilingual model. Lexical surprisal following improved the monolingual model by 19.5 percent and the bilingual by 17 percent. Note that "following" for WWD refers to the word that includes the disfluency. For example, if a syllable was repeated on a word, surprisal following refers to the word that has the syllable repetition. As seen in Table 3, preceding word integration cost (IC) was significant in predicting WWD in monolinguals well as in bilingual speakers, with words before disfluent words having lower IC. Finally, IC following was significant in the bilingual model, with disfluent words having lower integration costs, possibly to compensate for high surprisal.

3.4. Repetitions

Surprisal following was a significant predictor of repetitions in the monolingual as well as well as the bilingual model. Also, as shown in Table 4, we found storage cost (SC) following to be a significant feature both in the monolingual and bilingual model. In both cases, the words that were repeated exhibited higher SC. SC following improved the monolingual model AUC by 16 percent, and the bilingual model AUC by 7 percent. Lastly, we found that the duration of the repeated word was significant in predicting repetitions in bilinguals, with repeated words having shorter duration.

4. Conclusions

We found that all types of disfluencies tend to happen at high surprisal words, validating surprisal theory for both monolingual and bilingual speakers. We observed some interesting differences in disfluency manifestation between bilingual and monolingual. Specifically, we found evidence that bilingual speakers were cognitively fatigued after accessing and producing fluently a difficult word (e.g., low frequency and low neighborhood density), which left them with limited resources to plan the remaining of the utterance. In order to compensate for this, bilinguals chose phonetically simpler words, such as stop words with low IPC score and high bigram phonotactic probability, which are easier to construct, allowing them to better access upcoming words and maintain fluent speech. Monolinguals did not show this compensatory strategy. In their case, between-word disfluencies were more likely to occur before a difficult, low neighborhood density word, suggesting difficulties in lexical access.

Table 2. Comparison of features for Monolinguals and Bilinguals, Between-word Disfluencies. MN p-val column denotes McNemar's p-value. Statistically significant coefficients have p-value < 0.005.

	Monolinguals				Bilinguals			
Feature	**Coeff.**	**Std Err.**	**MN p-val**	**AUC**	**Coeff.**	**Std Err.**	**MN p-val**	**AUC**
Surprisal Prec.	0.0833	0.0207	-	0.5319	0.2564	0.0120	-	0.5764
Surprisal Fol.	**0.5953**	**0.0164**	**5.42E-57**	**0.6853**	**0.4896**	**0.0141**	**5.71E-23**	**0.6598**
IC Preceding	-0.1711	0.0092	1.69E-01	0.6879	-0.1259	0.0353	2.43E-01	0.6608
IC Following	0.2818	0.0467	3.15E-01	0.7020	0.1821	0.0237	2.93E-01	0.6622
SC Preceding	-0.0938	0.0246	1.87E-01	0.7039	-0.3630	0.0270	3.45E-01	0.6672
SC Following	0.0987	0.0201	9.79E-02	0.7057	0.1866	0.0209	2.73E-01	0.6790
Duration Prec.	0.1599	0.0159	2.96E-01	0.7127	0.0903	0.0153	2.86E-01	0.6810
Duration Fol.	-0.2568	0.0105	9.35E-02	0.7469	-0.3113	0.0073	1.54E-01	0.6855

Table 3. Comparison of features for Monolinguals and Bilinguals, Within-word Disfluencies. MN p-val column denotes McNemar's p-value. Statistically significant coefficients have p-value < 0.005.

	Monolinguals				Bilinguals			
Feature	**Coeff.**	**Std Err.**	**MN p-val**	**AUC**	**Coeff.**	**Std Err.**	**MN p-val**	**AUC**
Surprisal Prec.	0.0206	0.0065	-	0.4994	-0.2174	0.0148	-	0.5400
Surprisal Fol.	**0.5299**	**0.1676**	**1.00E-64**	**0.6943**	**0.7449**	**0.0137**	**1.35E-28**	**0.7097**
IC Preceding	**-0.5724**	**-0.1810**	**1.39E-04**	**0.7018**	**-0.3878**	**0.0151**	**6.61E-18**	**0.7181**
IC Following	-0.0728	-0.0230	4.25E-02	0.7028	**-0.5406**	**0.0219**	**1.24E-06**	**0.7225**
SC Preceding	0.3691	0.1167	1.30E-02	0.7032	0.6024	0.0163	1.39E-02	0.7468
SC Following	0.0482	0.0152	6.36E-02	0.7096	-0.2342	0.0155	9.89E-02	0.7579
Duration Prec.	0.0879	0.0278	9.09E-02	0.7114	0.0226	0.0146	3.08E-02	0.7584
Duration Fol.	-0.2519	-0.0797	6.06E-02	0.7145	-0.2830	0.0162	6.89E-02	0.7608

Table 4. Comparison of features for Monolinguals and Bilinguals, Word Repetitions. MN p-val column denotes McNemar's p-value. Statistically significant coefficients have p-value < 0.005.

	Monolinguals				Bilinguals			
Feature	**Coeff.**	**Std Err.**	**MN p-val**	**AUC**	**Coeff.**	**Std Err.**	**MN p-val**	**AUC**
Surprisal Prec.	0.4180	0.0200	-	0.6127	0.2505	0.0200	-	0.5876
Surprisal Fol.	**0.3890**	**0.0074**	**6.43E-5**	**0.6371**	**0.2646**	**0.0118**	**2.54E-06**	**0.5877**
IC Preceding	0.0369	0.0238	1.03E-01	0.6597	0.0760	0.0104	2.10E-01	0.5989
IC Following	0.1392	0.0202	3.26E-02	0.6634	0.4251	0.0316	3.17E-01	0.5978
SC Preceding	-1.0165	0.0342	3.26E-02	0.6907	-0.9147	0.0224	3.86E-01	0.6263
SC Following	**0.6643**	**0.0132**	**5.99E-03**	**0.7526**	**0.7749**	**0.0147**	**4.38E-14**	**0.6947**
Duration Prec.	0.0014	0.0137	7.73E-02	0.7532	-0.0712	0.0183	1.04E-01	0.6953
Duration Fol.	-0.2679	0.0129	3.75E-02	0.7553	**-0.3007**	**0.0083**	**4.67E-06**	**0.6993**

References

Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, 60(1):92–111, 2009.

Boerma, T. and Blom, E. Assessment of bilingual children: What if testing both languages is not possible? *Journal of Communication Disorders*, 66:65–76, 2017.

Brysbaert, M., Mandera, P., and Keuleers, E. The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, 27(1):45–50, 2018.

Byrd, C., Haque, A., and Johnson, K. Speech-language pathologists' perception of bilingualism as a risk factor for stuttering. *Journal of Communication Disorders, Deaf Studies & Hearing Aids*, 4(2):1000158, 2016.

Byrd, C. T. Assessing bilingual children: Are their disfluencies indicative of stuttering or the by-product of navigating two languages? In *Seminars in Speech and Language*, volume 39, pp. 324–332. Thieme Medical Publishers, 2018.

Dammalapati, S., Rajkumar, R., and Agarwal, S. Effects of duration, locality, and surprisal in speech disfluency prediction in english spontaneous speech. In *Proceedings of the Society for Computation in Linguistics 2021*, pp. 91–101, 2021.

De Lamo White, C. and Jin, L. Evaluation of speech and language assessment approaches with bilingual children. *International Journal of Language & Communication Disorders*, 46(6):613–627, 2011.

Fabiano-Smith, L. and Hoffman, K. Diagnostic accuracy of traditional measures of phonological ability for bilingual preschoolers and kindergarteners. *Language, Speech, and Hearing Services in Schools*, 49(1):121–134, 2018.

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

Grasso, S. M., Wagner Rodríguez, C. A., Montagut Colomer, N., Marqués Kiderle, S.-K., Sánchez-Valle, R., and Santos Santos, M. Á. Bilingual primary progressive aphasia: A scoping review of assessment and treatment practices. *Journal of Alzheimer's Disease*, pp. 1–24, 2023.

Grimm, A. and Schulz, P. Specific language impairment and early second language acquisition: The risk of over-and underdiagnosis. *Child Indicators Research*, 7:821–841, 2014.

Hale, J. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*, 2001.

Hemsley, G., Holm, A., and Dodd, B. Identifying language difference versus disorder in bilingual children. *Speech, Language and Hearing*, 17(2):101–115, 2014.

Jakielski, K. J. *Motor organization in the acquisition of consonant clusters*. The University of Texas at Austin, 1998.

Jusczyk, P. W., Luce, P. A., and Charles-Luce, J. Infants' sensitivity to phonotactic patterns in the native language. *Journal of memory and Language*, 33(5):630–645, 1994.

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Vitevitch, M. S. and Luce, P. A. A web-based interface to calculate phonotactic probability for words and nonwords in english. *Behavior Research Methods, Instruments, & Computers*, 36(3):481–487, 2004.

A. Additional Results

A.0.1. Additional Features for Between-word Disfluencies

According to Table 5, only neighborhood density following was a significant predictor of BWD in monolinguals, with words following disfluencies having fewer phonological neighbors. An opposite pattern was found in bilinguals, with words preceding disfluencies being lexically more difficult whereas words after disfluencies being phonetically easier. IPC following as well as bigram phonotactic probability following were significant predictors of BWD, with words following disfluencies being more likely to have lower IPC score and higher bigram phonotactic probability. Furthermore, significant predictors for bilinguals were word frequency preceding and neighborhood density preceding, with words preceding disfluencies being less frequent and having fewer phonological neighbors. Lastly, content word following was significant in the bilingual model, with words following disfluencies being less likely to be content words. The above findings can be illustrated through the following example: "give me the stethoscope uhm to hear your heart". In this example, "stethoscope" is the preceding word and "to" is the following word. We found that for bilinguals, the word before the disfluency (i.e., stethoscope) was more likely to have low frequency and be from a sparse neighborhood and the word after the disfluency (i.e., "to") was more likely to have low IPC score (i.e., easier in terms of phonological complexity), high bigram phonotactic probability (i.e., frequent sound sequence), and less likely to be a content word. In contrast, disfluencies tended to be before difficult words (i.e., low neighborhood density) in monolinguals (e.g., "give me the uhm stethoscope to hear your heart").

Table 5. Feature significance in predicting Between-word disfluencies. The significant features for bilinguals are IPC following (P<0.005), bigram phonotactic probability following (P<0.005), word frequency(P<0.005), neighbor density (P<0.05), content word following (P<0.005). The only significant feature for monolinguals was neighborhood density following (P<0.005).

	Monolinguals			**Bilinguals**		
Feature	**Coefficient**	**Std Error**	**p-val**	**Coefficient**	**Std Error**	**p-val**
IPC Preceding	0.0187	0.030	12.944184	0.0568	0.029	1.507
IPC Following	0.0269	0.037	11.397432	**-0.1841**	**0.033**	**0.001**
Bigrams Phono. Prob. Prec.	0.0439	0.021	1.106232	0.0132	0.033	16.647
Bigrams Phono. Prob. Fol.	0.1083	0.036	0.166344	**0.3707**	**0.029**	**<0.0001**
Word Frequency Preceding	-0.0904	0.027	0.08916	**-0.2026**	**0.024**	**<0.0001**
Word Frequency Fol.	-0.0357	0.031	6.288264	-0.0009	0.030	23.471
Neighbor Density Preceding	-0.0258	0.035	11.32584	**-0.1044**	**0.029**	**0.041**
Neighbor Density Fol.	**-0.1456**	**0.027**	**0.000816**	-0.0959	0.048	1.415
Content Word Preceding	0.0122	0.037	17.94276	0.0113	0.040	18.657
Content Word Following	-0.0306	0.035	9.31848	**-0.2593**	**0.039**	**<0.0001**
Stop Word Preceding	-0.0023	0.038	22.86672	-0.0721	0.040	2.022
Stop Word Following	-0.0750	0.030	0.508728	-0.0528	0.032	2.652

A.0.2. Additional Features for Within-word Disfluencies

According to Table 6, the only significant feature for WWD in the monolingual model was whether the disfluent word was a stop word; that is, disfluencies were less likely to occur on stop words. In the bilingual model, IPC of the word prior to the disfluent word, word frequency of disfluent word and whether the disfluent word was a stop word were all significant at predicting WWD. In particular, words before disfluent words were more likely to have lower IPC scores and disfluent words were less frequent and less likely to be stop words. These findings can be illustrated in the following example: "give me the ssstethoscope to hear your heart", the word stethoscope is the disfluent word and it contains a sound repetition. According to our monolingual model, the disfluent word (i.e., stethoscope) was less likely to be a stop word. In our bilingual model, we found that the word prior to the disfluent word (in the example that would be the word "the") was more likely to have low IPC score, that is, be phonetically less complex and easier. Also, the disfluent word (i.e., stethoscope) was more likely to be a low frequency word and less likely to be a stop word.

A.0.3. Additional Features for Repetitions

According to Table 7, word frequency preceding and content word following were significant features in the monolingual model. That is, words before a repeated word were more likely to have low frequency and repeated words were less likely to be content words. In the bilingual model, IPC of the repeated word (i.e., following), frequency of the repeated word,

Table 6. Feature significance in predicting Within-word disfluencies. The significant features for bilinguals are IPC (P<0.05), word frequency following (P<0.05), and stop word following (P<0.05). The significant feature for monolinguals is stop word following (P<0.005).

	Monolinguals			**Bilinguals**		
Feature	**Coefficient**	**Std Error**	**p-val**	**Coefficient**	**Std Error**	**p-val**
IPC Preceding	-0.0417	0.0326	5.5974	**-0.1027**	**0.0213**	**0.0230**
IPC Following	0.1500	0.0486	0.3122	0.0614	0.0415	4.1609
Bigrams Phono. Prob. Prec.	-0.0743	0.0184	0.0704	-0.0622	0.0404	3.8021
Bigrams Phono. Prob. Fol.	0.0085	0.0296	18.7489	-0.0469	-0.0148	0.2762
Word Frequency Prec.	0.1089	0.0484	1.2269	0.0889	0.0610	4.2938
Word Frequency Fol.	-0.0501	0.0238	1.5477	**-0.0969**	**0.0172**	**0.0076**
Neighbor Density Prec.	-0.0227	0.0237	8.7486	0.0056	0.0370	21.2196
Neighbor Density Fol.	-0.0628	0.0385	3.2908	-0.0062	0.0403	21.1800
Content Word Preceding	-0.0650	0.0317	1.6992	-0.1004	0.0318	0.2785
Content Word Following	0.0862	0.0244	0.1534	0.1434	0.0377	0.1003
Stop Word Preceding	0.0306	0.0377	10.5071	0.0569	0.0194	0.4018
Stop Word Following	**-0.2019**	**0.0324**	**0.0036**	**-0.1782**	**0.0369**	**0.0223**

whether the repeated word was content or stop, as well as frequency of the word before the repeated word were all significant predictors of repetitions. In the example "give me the stethoscope to to hear your heart", the word "stethoscope" is the preceding word and the word "to" is the following word (i.e., the word that is repeated). Based on both our models, the preceding word "stethoscope" was less likely to be high frequency word and the repeated word was less likely to be content word. Furthermore, in bilinguals, the repeated word (i.e., "to") was more likely to have low IPC score, more likely to be a high frequency word, and more likely to be a stop word.

Table 7. Feature significance in predicting repetitions. The significant features for bilinguals are word frequency (P<0.05), following word frequency (P<0.05), following content word (P<0.005), following stop word (P<0.005), and following word IPC (P<0.005). The significant feature for monolinguals are word frequency (P<0.05) and following content word (P<0.05).

	Monolinguals			**Bilinguals**		
Feature	**Coefficient**	**Std Error**	**p-val**	**Coefficient**	**Std Error**	**p-val**
IPC Preceding	-0.0385	0.0186	1.6541	0.0642	0.0286	1.2416
IPC Following	-0.0784	0.0188	0.0586	**-0.1727**	**0.0240**	**0.0012**
Bigrams Phono. Prob.Prec.	-0.0395	0.0312	5.6870	0.0116	0.0275	16.3919
Bigrams Phono. Prob. Fol.	-0.0675	0.0314	1.4407	-0.0182	0.0594	18.3976
Word Frequency Prec.	**-0.0453**	**0.0100**	**0.0352**	**-0.1250**	**0.0284**	**0.0411**
Word Frequency Fol.	0.0720	0.0328	1.3454	**0.2272**	**0.0488**	**0.0286**
Neighbor Density Prec.	0.0171	0.0258	12.5879	-0.0050	0.0330	21.1800
Neighbor Density Fol.	0.0020	0.0327	22.8442	0.0631	0.0312	1.7669
Content Word Prec.	0.0026	0.0289	22.3114	0.1094	0.0386	0.4701
Content Word Following	**-0.0899**	**0.0129**	**0.0016**	**-0.2155**	**0.0322**	**0.0022**
Stop Word Preceding	0.0040	0.0227	20.7367	-0.0950	0.0406	1.0565
Stop Word Following	0.0478	0.0233	1.7016	**0.1808**	**0.0208**	**0.0003**