Incremental Disfluency Detection for Spoken Learner English

Lucy Skidmore and **Roger K. Moore**
Speech and Hearing Research Group
Department of Computer Science
University of Sheffield, UK
{lskidmore1,r.k.moore}@shef.ac.uk

Abstract

Incremental disfluency detection provides a framework for computing communicative meaning from hesitations, repetitions and false starts commonly found in speech. One application of this area of research is in dialogue-based computer-assisted language learning (CALL), where detecting learners' production issues word-by-word can facilitate timely and pedagogically driven responses from an automated system. Existing research on disfluency detection in learner speech focuses on disfluency removal for subsequent downstream tasks, processing whole utterances non-incrementally. This paper instead explores the application of laughter as a feature for incremental disfluency detection and shows that when combined with silence, these features reduce the impact of learner errors on model precision as well as lead to an overall improvement of model performance. This work adds to the growing body of research incorporating laughter as a feature for dialogue processing tasks and provides further support for the application of multimodality in dialogue-based CALL systems.

1 Introduction

Speech disfluencies such as hesitations, repetitions and false starts are an inherent artefact of spoken language. Systematic in their structure, disfluencies comprise of a reparandum phrase, optional interregnum phrase and repair phrase (Levelt, 1983).

I'd like a [coffee + {uh} tea] please
(coffee: reparandum; {uh}: interregnum; tea: repair)

Following the notation scheme described by Shriberg (1994), the example above shows the components of a disfluency. The speaker changes their drink order by replacing "*coffee*" with the intended word "*tea*". The + represents the 'interruption point', often marked prosodically with features such as silence or reparandum word cutoff. The following, optional interregnum phrase can contain filled pauses such as "*uh*" like in the example, edit terms such as "*I mean*" and finally discourse markers such as "*you know*".

Detecting such disfluencies is of particular interest in the context of dialogue-based CALL, where learners interact with an automated system in order to practice conversation in the language that they are learning. In the task-based scenario where a learner is practicing ordering a drink at a café, having a system that can identify and appropriately respond to learners' disfluencies in real-time is highly desirable and not something that is available in existing approaches thanks to the lack of incremental processing (Bibauw et al., 2019).

With the above in mind, this work builds on incremental disfluency detection research and applies it to a language learning setting. The nature of disfluencies in learner speech are explored and learner errors are identified as an area of difficulty in existing approaches. Subsequently, the non-lexical features of laughter and silence are suggested as possible solutions to this issue and their impact is tested and compared to a baseline model. Findings are reported and considerations for future work in this area are discussed.

2 Related Work

Disfluency detection is a widely studied area of research, with the most successful approaches leveraging BERT transformer models to achieve high accuracy (e.g. Bach and Huang, 2019; Jamshid Lou and Johnson, 2020; Rocholl et al., 2021). These models operate non-incrementally using whole sentences as inputs, often with a view to remove the disfluencies from transcripts all together.

This is also the case for research on disfluency detection in learner speech, which has been applied to improve the downstream tasks of grammatical error detection and correction using bi-directional LSTMs (Lu et al., 2019) as well as end-to-end mod-

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 272 - 278

els (Lu et al., 2020). Approached as a sequence labelling task, disfluencies are flattened and models are trained to detect the reparandum phrase. This approach is not suited to spoken dialogue systems, however, which benefit from word-by-word processing and the retention of all parts of the disfluency in order to generate meaningful and timely responses (Schlangen and Skantze, 2009). In a language learning context, such capabilities would not only enable conversational systems to better employ incremental feedback strategies such as prompting but also provide insight into the nature of individual learners' disfluency behaviours.

Incremental disfluency detection addresses the issues described above and forms a smaller subsection of research. Restricted by their left-to-right operability, incremental systems detect disfluency at the point of repair onset and subsequently 'look-back' for the reparandum phrase. To date, there has only been one research paper related to incremental disfluency detection for learner speech, where Moore et al. (2015) reported poor performance when using an incremental dependency parser trained on native data. Various approaches have been tested using the Switchboard Corpus (Godfrey et al., 1992), however. These are described below.

Following a noisy channel approach, Hough and Purver (2014) implemented a pipeline of Random Forest classifiers detecting interregna, repair and reparandum phrases separately using input features derived from trigram language models for words and POS tags. Simplifying the task to a one model, multi-class sequence labelling problem using deep neural networks, Hough and Schlangen (2015) successfully applied a RNN using only word embeddings and POS tags as input features. This approach was extended further, using LSTMs for the joint tasks of utterance segmentation (Hough and Schlangen, 2017) as well as multi-task learning with utterance segmentation, POS tagging and language modelling (Rohanian and Hough, 2020). Current state-of-the-art performance is achieved by Rohanian and Hough (2021), who incrementalised a BERT-based disfluency detector by using utterance predictions from a GPT-2 language model as inputs to the model.

With the exception of word timings (Hough and Schlangen, 2017; Rohanian and Hough, 2020, 2021) the incremental approaches outlined above have yet to explore the impact of non-lexical features on disfluency detection, despite having been successfully integrated into non-incremental settings (Zayats et al., 2016; Lu et al., 2020). Considering the fact that incremental detection begins at repair onset, it seems likely that leveraging paralinguistic information associated with the interruption point will be beneficial to detection. Approaches to such integration are explored in this work.

3 Disfluencies in Learner Speech

On average, disfluencies occur at a higher rate in learner speech compared to native speech (Hilton, 2008; De Jong et al., 2013). Learner speech disfluency datasets also contain longer reparandum phrases compared to native equivalents (Lu et al., 2020). This is in part thanks to language learners having a lower degree of 'automatisation' in the language they are learning (Temple, 1992) and is cited by Moore et al. (2015) as the reason why disfluencies in learner speech are more difficult to detect automatically.

Another artefact of learner speech disfluencies is their co-occurrence with learner errors. The examples below highlight how errors interact with disfluencies in the NICT-JLE Corpus used for this study. The disfluency phrases are labelled and words in bold indicate learner errors.

(1) My computer [**use** + {er} is used] by [**all family** + my family]

(2) She [[**wanted shopping** + **wanted shop**] + {er} wanted to go shopping]

(3) [[I don't + **I'm not have watching movie**] + I don't have **no** time to **watch movie**]

As the examples show, learner errors can occur in the reparandum phrase, the repair phrase, or both. The first example shows an instance where the learner error occurs in the reparandum phrase and is then subsequently repaired to its correct form. The second example shows how this can occur in a nested disfluency, where the inner disfluency instance contains learner errors in both the reparandum and repair phrases, with the outer disfluency instance being without error. The third example shows an instance where the initial reparandum phrase is correct but the subsequent repair phrases both contain errors.

The presence of learner errors is often cited as a contributing factor to the difficulty of other NLP tasks for learner language data such as parsing

(Napoles et al., 2016) and POS tagging (Nagata et al., 2018). With this in mind, it was hypothesised that learner errors would have a similar negative effect on disfluency detection and so their impact was tested as part of this experimentation.

4 Silence and Laughter

Incorporating instances of silence is a successful method of increasing model performance in non-incremental disfluency detection research. Silence has been encoded explicitly using its presence or absence as an input feature (Liu et al., 2005; Ferguson et al., 2015), implicitly through the inclusion of audio features such as filter banks (Lu et al., 2020) and even as a prediction of prosodic cues from text (Zayats and Ostendorf, 2019). Research into the nature of silence in learner speech has shown that non-native speakers are more likely to pause mid-clause than native speakers during linguistic processes such as repair (Tavakoli, 2011). With this in mind, it seems likely that including silence features will have a positive impact on the model performance and so is tested here.

Language learners use laughter as a 'trouble management device' during uncertainty (Looney and He, 2021), when pre-empting a problematic action (Petitjean and González-Martínez, 2015) and after making an error (Gao and Wu, 2018). In an analysis of UK university English proficiency interviews of 23 Chinese students, Gao (2020) found that laughter co-occurs with disfluencies in three ways: *(i)*, on its own between the reparandum and repair phrase, *(ii)*, alongside indicators of an interruption point such as pauses and word cutoffs, and *(iii)*, simultaneously as laughed speech either during the repair phrase or the whole disfluency. Laughter has been shown to improve performance of models for other dialogue processing tasks such as dialogue act classification (Maraev et al., 2021) but as of yet, has not been applied as a feature to detect disfluencies in learner speech.

5 Experimentation Set Up

5.1 NICT-JLE Corpus

The National Institute of Information and Communications Technology Japanese Learner English (NICT-JLE) Corpus is a transcription-only corpus of 1,281 English oral proficiency tests (approximately 300 hours of speech) of Japanese speaking learners of English (Izumi et al., 2004). The test, known as the Standard Speaking Test (SST) is carried out in an interview style between learner and assessor, where the learner is asked to perform a selection of various tasks. These include engaging in open dialogue, a role-play scenario and a picture description task. Each transcribed interview contains labels for edit terms and disfluencies, 'non-verbal sounds' (including silence and laughter), as well as meta-data such as the learners' SST level, gender and nationality. 167 of the interviews contain additional labels for learners' morphological, grammatical and lexical errors.

total words	1,165,785
disfluency instances per 100 words	7.54
edit terms per 100 words	11.55
learner errors per 100 words	11.10

Table 1: Dataset statistics for the NICT-JLE Corpus.

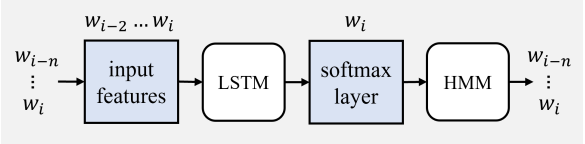

Figure 1: Diagram of the model structure used for experimentation.

For experimentation the corpus was lemmatized using the NLTK WordNet Lemmatizer (Bird et al., 2009) and POS-tagged by the Stanford POS-tagger (Toutanova et al., 2003). Learner utterances (excluding those that contained Japanese) were extracted from the transcripts and split with 80% of the data for training, 10% for heldout and 10% for testing, ensuring that each dataset had an equal distribution of SST levels and that all transcripts in the test set were taken from the subset that contained tagged learner errors. Dataset statistics are summarised in Table 1[1]. Note that the figure for learner error rates reflects the test set only.

5.2 Model

Following Hough and Schlangen (2017), the approach used for experimentation combines an LSTM model with an HMM decoder. As visualised in Figure 1, the model processes sequences incrementally in a maximum window of nine words to accommodate the repair start and the eight words prior. Features are extracted from the trigram $w_{i-2}...w_i$ and used as inputs to the LSTM. The

[1]GitHub repository of adapted dataset: https://github.com/lucyskidmore/nict-jle

Model	F_{rpS}	F_{rm}	F_e
baseline	0.757	0.723	**0.982**
+ silence	0.759	0.726	0.981
+ laughter	0.754	0.719	**0.982**
+ silence and laughter	**0.766**	**0.732**	**0.982**

Table 2: F-score results of the baseline compared to silence and laughter models for repair start, reparandum phrase and edit term detection.

LSTM network contains a hidden layer of 50 nodes and an output layer of size ten, reflecting the size of the tag set. Negative log likelihood is used as the cost function, as well as stochastic gradient descent over the parameters, including the word embeddings. The learning rate is 0.005 and L2 regularisation is applied to the parameters with a weight of 0.0001. The LSTM softmax output layer is used as an input to the HMM where outputs are updated incrementally with the best sequence hypothesis from Viterbi decoding.

5.3 Input Features

The baseline model uses trigrams of POS tags and fastText word embeddings (Bojanowski et al., 2017) of size 50 as input features. Silence and laughter features were derived directly from the NICT-JLE transcripts. Each word was assigned a vector, indicating the presence or absence of a preceding short pause, long pause, laughter, or if the word itself was laughed for the current word and previous two words.

5.4 Disfluency Tags

Following Hough and Schlangen (2015), disfluencies are labelled at repair onset as `rpS-n` as illustrated below, where `n` denotes the distance to the reparandum start from the repair start.

I'd	like	a	[coffee	{uh}	tea]	please
f	f	f	f	e	rpS-2	f

This approach allows for both incrementality and the labelling of nested disfluencies. Edit terms are combined with interregna and labelled as `e` and 'fluent' words are labelled as `f`. The maximum length of a disfluency is cut off at `rpS-8` which results in a total tag set size of ten.

6 Results

Table 2 reports the F-score results of the baseline model compared to the models with additional non-lexical features. F-scores are reported for repair start as well as reparandum phrase (commonly used to measure non-incremental performance) and edit term detection. Despite individually having little impact on baseline performance, when combined, the features of silence and laughter lead to an improvement in both repair start and reparandum phrase detection. Edit term detection performance remains high across all model variations.

Error Pos.	Model	P_{rpS}	R_{rpS}	F_{rpS}
rpS	baseline	0.730	0.757	0.744
	+ S&L	0.749	0.759	0.754
rpS-1	baseline	0.398	0.713	0.511
	+ S&L	0.432	0.726	0.541
rpS, *rpS-1*	baseline	0.481	0.768	0.592
	+ S&L	0.518	0.773	0.621

Table 3: Precision, recall and F-score results for repair start detection of disfluency phrases with co-occurring learner errors.

Model	Inc.?	Corpus	F_{rm}
+ S&L	✓	NICT-JLE	0.732
Moore et al. (2015)	✓	BULATS	0.478
Lu et al. (2019)	-	NICT-JLE	0.798

Table 4: Reparandum phrase F-score results of the final model compared to existing approaches with varying corpora and incrementality.

Table 3 reports the precision, recall and F-score results for repair start detection of disfluencies that co-occur with learner errors. Reflecting the three scenarios described in Section 3, disfluencies that co-occur with an error at repair onset (*rpS*), an error immediately preceding the repair phrase start (*rpS-1*) and errors occurring both immediately before and at repair onset (*rpS-1* and *rpS*) are reported. Firstly, comparing the baseline performance of all three scenarios with the overall baseline performance reported in Table 2 reveals the extent to which learner errors impact model performance — this is especially true for disfluency instances that are preceded by a learner error. In turn, it is these instances that show the most improvement in performance when silence and laughter features are included, with precision being particularly boosted.

Table 4 compares the performance of the adapted model with two existing approaches to disfluency detection in learner speech: an incremental model

tested on the BULATS Corpus[2] (Moore et al., 2015) and a non-incremental model tested on the NICT-JLE Corpus (Lu et al., 2019). Neither approach reports repair start detection so only reparandum phrase detection is compared here. Although not directly comparable due to the mismatches in corpora and incrementality, the results from this paper significantly outperform Moore et al. (2015), setting a new benchmark for incremental disfluency detection for learner speech. As expected, performance does not reach the level of current state-of-the-art non-incremental approaches.

7 Discussion

The results from this experimentation give support to the integration of paralinguistic features for incremental disfluency detection in learner speech. The impact of silence and laughter on the detection precision of disfluencies that co-occur with learner errors highlights the value of such features in settings where lexical data is 'non-typical'. This is of particular importance in incremental approaches where detection occurs at repair onset, with a reduced reliance on the syntactic parallelism between reparandum phrase and repair phrase often exploited by non-incremental systems.

Despite the improvements described above, overall performance gains are small and remain lower than non-incremental approaches. However, there are further approaches to model improvement worth exploring. Firstly, following the recent work of Rohanian and Hough (2021), it would be of interest to test the impact of an incrementalised BERT-based detector on learner speech. Secondly, using a POS-tagger specifically for learner speech such as that developed by Nagata et al. (2018) may help boost performance. It would also be beneficial to investigate the impact of these adaptations on other aspects of learner speech that inform disfluency behaviour, including learners' first language, task type and proficiency level.

Another limitation of the study is that the NICT-JLE Corpus is a transcription-only dataset with limited features. Without audio files available, instances of silence and laughter are derived directly from transcripts. In the same way that ASR output deteriorates disfluency detection performance compared to transcribed data (Lu et al., 2019), it is likely that automatic laughter and silence detection derived from audio would have a similar effect and may not be as impactful for model improvement. In addition, it would be interesting to investigate the relationship between learner errors and disfluencies by modelling these features jointly. However, in the NICT-JLE Corpus, learner error tags are only available for the test set and so cannot be used as features in training. Furthermore, the performance boost shown when combining laughter together with silence provides the motivation to explore additional paralinguistic features in combination, such as gestures and gaze, both of which have been shown to be used in conversation to signal disfluency (Chen et al., 2002; Radford, 2009). Finally, as the NICT-JLE Corpus is a collection of assessor-learner conversations, it is not clear if learners would still enact the same strategies of laughter to indicate disfluencies when practising with a dialogue-based CALL system.

8 Future Work

To the best of our knowledge, there is currently no publicly available resource that addresses the limitations of the NICT-JLE Corpus outlined above. With this in mind, there is a strong case to be made for the development of a multimodal corpus for use in dialogue-based CALL applications, collected by means of a 'Wizard of Oz' experiment with language learners and human language tutors. Audio, video and transcript files annotated with disfluencies, edit terms, learner errors as well as paralinguistic information would provide ample opportunity for research into both incremental disfluency detection and also other dialogue processing tasks.

9 Conclusion

In conclusion, this work tested the impact of laughter and silence as features for incremental disfluency detection of learner speech. When combined, these features show an overall improvement in model performance, increasing precision for disfluencies that co-occur with learner errors. To date, this is the first work to use laughter as a feature for disfluency detection in a language learning setting, with the resulting model significantly outperforming previous incremental approaches for learner speech. These findings act as a starting point for the further integration of paralinguistic features for incremental disfluency detection and help make the case for the development of a multimodal corpora for dialogue-based CALL applications.

[2]This corpus was provided to the researchers by Cambridge Assessment English and is not publicly available.

References

Nguyen Bach and Fei Huang. 2019. Noisy BiLSTM-Based Models for Disfluency Detection. In *Proceedings of Interspeech 2019*, pages 4230–4234.

Serge Bibauw, Thomas François, and Piet Desmet. 2019. Discussing with a computer to practice a foreign language: research synthesis and conceptual framework of dialogue-based CALL. *Computer Assisted Language Learning*, pages 1–51.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Lei Chen, Mary Harper, and Francis Quek. 2002. Gesture patterns during speech repairs. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ICMI '02, page 155, USA. IEEE Computer Society.

Nivja H. De Jong, Margarita P. Steinel, Arjen Florijn, Rob Schoonen, and Jan H. Hulstijn. 2013. Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, 34(5):893–916.

James Ferguson, Greg Durrett, and Dan Klein. 2015. Disfluency detection with a semi-Markov model and prosodic features. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 257–262, Denver, Colorado. Association for Computational Linguistics.

Yan Gao. 2020. Laughter as Same-Turn Self-Repair Initiation in L2 Oral Proficiency Interview. *Open Journal of Social Sciences*, 8(4):479–494.

Yan Gao and Yaxin Wu. 2018. Laughter as Responses to Different Actions in L2 Oral Proficiency Interview. *Open Journal of Modern Linguistics*, 8(6):199–220.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520. IEEE Computer Society.

Heather Hilton. 2008. The link between vocabulary knowledge and spoken L2 fluency. *Language Learning Journal*, 36(2):153–166.

Julian Hough and Matthew Purver. 2014. Strongly incremental repair detection. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 78–89, Doha, Qatar. Association for Computational Linguistics.

Julian Hough and David Schlangen. 2015. Recurrent neural networks for incremental disfluency detection. In *Proceedings of Interspeech 2015*, pages 849–853.

Julian Hough and David Schlangen. 2017. Joint, incremental disfluency detection and utterance segmentation from speech. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 326–336, Valencia, Spain. Association for Computational Linguistics.

Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2004. The NICT JLE Corpus: Exploiting the language learners' speech database for research and education. *International Journal of the Computer, the Internet and Management*, 12(2):119–125.

Paria Jamshid Lou and Mark Johnson. 2020. Improving disfluency detection by self-training a self-attentive model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3754–3763, Online. Association for Computational Linguistics.

Willem JM Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

Yang Liu, Elizabeth Shriberg, Andreas Stolcke, and Mary Harper. 2005. Comparing HMM, maximum entropy, and conditional random fields for disfluency detection. In *Proceedings of Interspeech 2005*, pages 3313–3316.

Stephen Daniel Looney and Yingliang He. 2021. Laughter and smiling: sequential resources for managing delayed and disaligning responses. *Classroom Discourse*, 12(4):319–343.

Yiting Lu, Mark J. F. Gales, Katherine M. Knill, Potsawee Manakul, and Yu Wang. 2019. Disfluency Detection for Spoken Learner English. In *Proceedings of the 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, pages 74–78.

Yiting Lu, Mark J.F. Gales, and Yu Wang. 2020. Spoken Language 'Grammatical Error Correction'. In *Proceedings of Interspeech 2020*, pages 3840–3844.

Vladislav Maraev, Bill Noble, Chiara Mazzocconi, and Christine Howes. 2021. Dialogue act classification is a laughing matter. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue*, pages 120–131.

Russell Moore, Andrew Caines, Calbert Graham, and Paula Buttery. 2015. Incremental dependency parsing and disfluency detection in spoken learner english. In *International Conference on Text, Speech, and Dialogue*, pages 470–479. Springer.

Ryo Nagata, Tomoya Mizumoto, Yuta Kikuchi, Yoshifumi Kawasaki, and Kotaro Funakoshi. 2018. A POS tagging model adapted to learner English. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The*

4th Workshop on Noisy User-generated Text, pages 39–48, Brussels, Belgium. Association for Computational Linguistics.

Courtney Napoles, Aoife Cahill, and Nitin Madnani. 2016. The effect of multiple grammatical errors on processing non-native writing. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11, San Diego, CA. Association for Computational Linguistics.

Cécile Petitjean and Esther González-Martínez. 2015. Laughing and smiling to manage trouble in french-language classroom interaction. *Classroom Discourse*, 6(2):89–106.

Julie Radford. 2009. Word searches: on the use of verbal and non-verbal resources during classroom talk. *Clinical linguistics & phonetics*, 23(8):598–610.

Johann C. Rocholl, Vicky Zayats, Daniel D. Walker, Noah B. Murad, Aaron Schneider, and Daniel J. Liebling. 2021. Disfluency Detection with Unlabeled Data and Small BERT Models. In *Proceedings of Interspeech 2021*, pages 766–770.

Morteza Rohanian and Julian Hough. 2020. Re-framing incremental deep language models for dialogue processing with multi-task learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 497–507, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Morteza Rohanian and Julian Hough. 2021. Best of both worlds: Making high accuracy non-incremental transformer-based disfluency detection incremental. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3693–3703, Online. Association for Computational Linguistics.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 710–718, Athens, Greece. Association for Computational Linguistics.

Elizabeth Ellen Shriberg. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, University of California, Berkeley.

Parvaneh Tavakoli. 2011. Pausing patterns: Differences between L2 learners and native speakers. *ELT journal*, 65(1):71–79.

Liz Temple. 1992. Disfluencies in learner speech. *Australian Review of Applied Linguistics*, 15(2):29–44.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259.

Vicky Zayats and Mari Ostendorf. 2019. Giving attention to the unexpected: Using prosody innovations in disfluency detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 86–95, Minneapolis, Minnesota. Association for Computational Linguistics.

Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2016. Disfluency Detection Using a Bidirectional LSTM. In *Proc. Interspeech 2016*, pages 2523–2527.