Adapting the NICT-JLE Corpus for Disfluency Detection Models

Lucy Skidmore
University of Sheffield, UK
lskidmore1@shef.ac.uk

Roger K. Moore
University of Sheffield, UK
r.k.moore@shef.ac.uk

Abstract

The detection of disfluencies such as hesitations, repetitions and false starts commonly found in speech is a widely studied area of research. With a standardised process for evaluation using the Switchboard Corpus, model performance can be easily compared across approaches. This is not the case for disfluency detection research on learner speech, however, where such datasets have restricted access policies, making comparison and subsequent development of improved models more challenging. To address this issue, this paper describes the adaptation of the NICT-JLE corpus, containing approximately 300 hours of English learners' oral proficiency tests, to a format that is suitable for disfluency detection model training and evaluation. Points of difference between the NICT-JLE and Switchboard corpora are explored, followed by a detailed overview of adaptations to the tag set and meta-features of the NICT-JLE corpus. The result of this work provides a standardised train, heldout and test set for use in future research on disfluency detection for learner speech.

1 Introduction

Consider the utterance below:

$$\text{I'd like a } [\underbrace{\text{coffee}}_{\text{reparandum}} + \underbrace{\{\text{uh}\}}_{\text{interregnum}} \underbrace{\text{tea}}_{\text{repair}}] \text{ please}$$

The speaker initially asks for a coffee but changes their request to tea instead. The linguistic mechanism by which this is achieved is known as self-repair, or in computer science research, disfluency. Disfluencies are comprised of a reparandum phrase, optional interregnum phrase and repair phrase, often marked prosodically at the 'interruption point' + with features such as silence or reparandum word cutoff (Levelt, 1983). Interregna 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*".

There are two perspectives from which disfluency detection research is applied. The first is disfluency removal, where transcribed speech is transformed into a form more similar to written text for subsequent downstream NLP tasks. The second is incremental detection, whereby disfluent phrases are detected word-by-word and retained to infer meaning for use in spoken dialogue systems. The most successful approaches leverage BERT language models (Devlin et al., 2019) to achieve high accuracy for both non-incremental (Bach and Huang, 2019; Jamshid Lou and Johnson, 2020; Rocholl et al., 2021) and incremental (Rohanian and Hough, 2021) frameworks.

Disfluency detection has also been explored for learner speech, using parsing-based approaches (Moore et al., 2015), bi-directional LSTMs (Lu et al., 2019) and end-to-end models (Lu et al., 2020) for the downstream task of grammatical error correction. These approaches train models using native speech and evaluate using datasets with restricted access (Caines et al., 2017) or small subsets of publicly available data (Izumi et al., 2004). This approach is undesirable for two reasons: *(i)* model performance cannot easily be compared with other work due to the limited access to evaluation materials for the task, and *(ii)*, models trained on native data creates a performance bias towards higher proficiency learners (Moore et al., 2015).

With the above in mind, this work expands on that of Skidmore and Moore (2022), who trained and evaluated an incremental disfluency detection model using the NICT-JLE corpus. The corpus is introduced and its features are explored through a comparative analysis with the Switchboard corpus. This is followed by a detailed overview of the tag set adaptation, POS-tagging approach, additional meta-features and the train, heldout and test divisions. This paper concludes with a discussion of directions for future work as well as limitations of the adapted corpus.

Table 1: General linguistic and disfluency features of the NICT-JLE (NICT) and Switchboard (SWBD) corpora.

	NICT	SWBD
total words	1165785	746290
total utterances	178934	102169
vocabulary size	13499	16810
utterance length (SD)	6.51 (3.27)	7.30 (3.61)
disfluency/100 words	7.54	3.56
edit term/100 words	11.55	5.16
rm length (SD)	1.62 (1.08)	1.58 (1.12)
nested rate	39.35	21.68
non-repetitious rate	45.60	49.45
with-interregnum rate	31.09	22.17

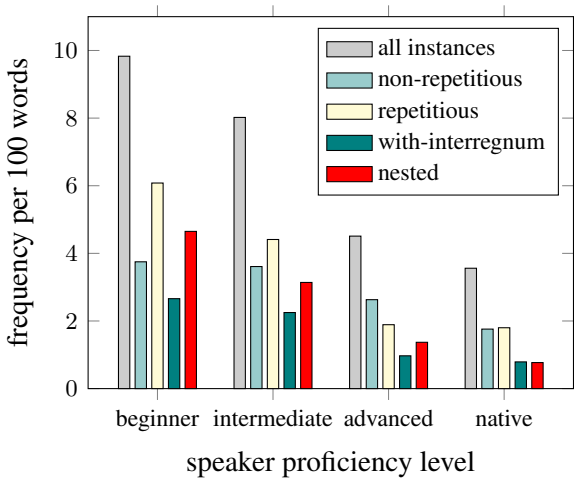

Figure 1: Frequency per 100 words of disfluency types according to speaker proficiency level.

2 The NICT-JLE Corpus

The National Institute of Information and Communications Technology Japanese Learner English (NICT-JLE) Corpus contains approximately 300 hours of transcribed oral proficiency tests of 1,281 Japanese-speaking learners of English (Izumi et al., 2004). The Standard Speaking Test (SST) is made up of three tasks for a learner to carry out with the assessor: open dialogue, a role-play scenario and a picture description task. Each transcribed test contains HTML-style tags for edit terms and disfluencies, 'non-verbal sounds' (including silence and laughter), as well as meta-data such as the learners' proficiency level, gender and nationality.

2.1 Disfluencies in the NICT-JLE corpus

Table 1 compares a range of general linguistic and disfluency features of the NICT-JLE corpus with the Switchboard corpus—the standard corpus used for disfluency detection on native speech. The NICT-JLE corpus is the larger of the two, with a higher number of both words and utterances. The Switchboard corpus has a larger vocabulary size and longer average utterance length. These figures are reflective of second language acquisition research that determines both vocabulary size and average utterance length as predictors of learners' speaking skills (Koizumi and In'nami, 2013; Hilton, 2008), where a larger vocabulary equates to a higher speaking proficiency.

Looking at the disfluency features, the NICT-JLE corpus has over twice as many disfluencies and edit terms per 100 words compared to the Switchboard corpus. These figures are again echoed in prior research, which attributes the lower degree of language 'automatisation' of language learners to the increased number of disfluencies found in learner speech (Wiese; Temple, 1992), with the same behaviour also seen for filled pauses (Hilton, 2008; De Jong et al., 2013). Reparandum phrase lengths, however, are comparable between the two corpora. Finally, when considering the features of the disfluencies themselves, the NICT-JLE corpus has higher rates of both nested and with-interregnum disfluencies, whereas the Switchboard corpus has a higher rate of non-repetitious disfluencies.

Exploring the relationship between disfluency behaviour and proficiency level further, Figure 1 shows the frequency per 100 words of all disfluency instances, and subsequently, repetitious, non-repetitious, with-interregnum and nested instances according to speaker proficiency level. The 'beginner', 'intermediate', and 'advanced' levels refer to speakers from the NICT-JLE corpus at SST levels one to three, four to six and seven to nine, respectively. The 'native' group refers to all speakers in the Switchboard corpus. As can be seen, the frequency of all disfluency types decreases as speaker proficiency level increases. Notably, the ratio of repetitious to non-repetitious disfluency instances is at its highest for beginner proficiencies and marginally higher for intermediate proficiencies. For advanced speakers, the distribution switches and non-repetitious disfluencies are shown to be more frequent. For native speakers, the distributions are approximately equal. These observations are again reflective of prior research,

which has shown the frequency of repetitions in learner speech to decline as linguistic knowledge increases (De Jong et al., 2013) as well as the distribution of features becoming more similar to that of a native speaker (Van Hest, 1996).

2.2 Learner errors in the NICT-JLE corpus

The NICT-JLE corpus also contains learner errors both inside and outside of disfluencies. The examples below illustrate instances of the former, with errors occurring in the reparandum phrase, the repair phrase, or both. 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**]

There are 167 interviews in the NICT-JLE corpus that contain additional labels for learners' morphological, grammatical and lexical errors. From this subset, there are approximately 11 instances of learner errors per 100 words. However, the actual rate of errors in the corpus is likely to be higher as errors that occur as part of a reparandum phrase are not annotated.

3 Adapting the NICT-JLE Corpus

The adapted version of the NICT-JLE corpus is available online[1]. With the focus of this task being learner speech, any utterances of assessors are omitted. In addition, utterances that contain Japanese, or partial sentences due to the retraction of personally identifiable details are removed. Individual file numbers are retained and utterance segmentation follows the original transcriptions.

3.1 Tokenization and POS tags

The corpus was tokenized using the Natural Language Toolkit (NLTK) (Bird et al., 2009) and subsequently tagged with parts-of-speech (POS) using Stanford's left3words MaxentTagger (Toutanova et al., 2003). Following the same conventions as the Switchboard corpus, contracted forms of words such as "I've" and "can't" were split by tokenization and re-merged after POS tag labelling to form compounded POS tags. In the case of the two examples, the POS tags for these words are `PRPVBP` and `MDRB`, respectively.

3.2 Disfluency tag sets

In order to match the conventions used in previous approaches to disfluency detection, the original disfluency labels in the NICT-JLE corpus were adapted to include two labelling styles: incremental tags (Hough and Schlangen, 2015) and 'beginning-inside-outside-end-single' (BIOES) tags (Lu et al., 2019, 2020), depicted in Figure 2. As can be seen, the incremental approach includes the repair phrase start in the tag set, with the relative position of the reparandum start denoted by `rps-n`. Here `f` refers to 'fluent' text and `i` refers to interregna. The BIOES tags only label the reparandum phrase, as 'single reparandum' `S-RM` for one word reparanda and as 'begin reparandum' `B-RM`, 'inside reparandum' `I-RM` or 'end of reparandum' `E-RM` for longer reparandum phrases. All other tokens are considered 'outside' `O` of the disfluent phrase. This approach removes the nesting structure from disfluencies to create one large disfluency instance. Interregna tags are omitted from the BIOES approach as they are often omitted from the tag set altogether during experimentation (Lu et al., 2019, 2020). The original labelling scheme for the NICT-JLE corpus does not include repair phrases so they are not included in the adapted tag sets.

3.3 Meta features

Alongside labels for disfluencies and learner errors, the NICT-JLE corpus also contains transcriptions for learners' 'non-verbal sounds'. From these annotations, silence and laughter were included as features in the adapted dataset due to their prevalence in the corpus as well as their value in previous disfluency detection (Ferguson et al., 2015; Lu et al., 2020) and dialogue processing (Maraev et al., 2021) applications. In the adapted corpus, each word is additionally assigned four binary values (1 or 0), indicating the presence or absence of a preceding short pause, long pause and laughter, as well as if the word itself was laughed. For the test set only, an additional binary value for each word is included to indicate the presence of a learner error. For 'omittance' type errors such as missed prepositions, the word preceding the omittance is labelled as erroneous. For example, in the utterance "I eat dinner in centre of Tokyo", 'in' would be labelled. Information regarding task type is also retained,

[1] https://github.com/lucyskidmore/nict-jle

	[I	can't	do +	{ah}	[I +	I]	couldn't	do]	my	work
Incremental:	f	f	f	i	rps-4	rps-1	f	f	f	f
BIOES:	B-RM	I-RM	I-RM	-	I-RM	E-RM	O	O	O	O

Figure 2: The incremental and BIOES tag sets for the adapted NICT-JLE corpus.

Table 2: General disfluency features for the train, heldout and test sets of the adapted NICT-JLE corpus.

	Train	Heldout	Test
disfluency/100 words	7.53	7.20	7.89
edit term/100 words	11.53	11.30	11.90
rm length	2.04	2.02	2.08

due to the proven influence of activity on leaner disfluency behaviour (Kormos, 1998). Each word is labelled as 'conversation', 'picture description' or 'role play', relating to the current activity type of the learner. For the latter two, the corpus also includes information regarding the topic of the activity. Finally, the speaker-level features of gender and English proficiency level (SST score) are also included.

3.4 Train, heldout and test datasets

Following the approach of the Switchboard corpus, the adapted NICT-JLE corpus was split with 80% of the files for training, 10% for heldout and 10% for testing. Each set has a near-equal distribution of learners according to English proficiency and all data for the test set has accompanying learner error tags. Table 2 shows the disfluency statistics for each of the split datasets, showing equivalency across sets.

4 Discussion

The comparative analysis of the NICT-JLE and Switchboard corpora reveals key differences in terms of linguistic complexity and disfluency behaviour. The native speech of the Switchboard corpus is more lexically complex, with longer utterances, wider vocabulary and a lower ratio of repetitious disfluencies. The disfluencies of the NICT-JLE corpus can be considered more 'stutter-like' with higher rates of edit terms, repetitions and learner errors. The parallelism between the data reported here and prior research on learner disfluency behaviour provides support for the adapted NICT-JLE corpus to be used as a proxy for learner speech more generally, highlighting key areas of attention for future research. Examples of such areas include the impact of nested disfluencies, edit terms and learner errors on detection accuracy.

The inclusion of various tag sets for incremental and non-incremental approaches provides ample opportunity for future research using the adapted NICT-JLE corpus. The most pertinent of which would be to apply the current state-of-the-art approaches using BERT language models (Bach and Huang, 2019; Rohanian and Hough, 2021) to a language learning setting, using the adapted NICT-JLE corpus for fine-tuning. Furthermore, the inclusion of prosodic and speaker-level features not only allows for multimodal models to be developed, such as those introduced by Skidmore and Moore (2022), but also provides a framework for more in-depth model analysis from a language learning perspective, such as the impact of activity type and speaking proficiency on model performance. Finally, the adapted corpus has the potential for further development. One example would be the inclusion of language assessors' utterances in order to accommodate other dialogue-processing tasks such as end-of-turn prediction and utterance segmentation.

5 Summary

In summary, this work details the adaptation of the NICT-JLE corpus to be used as training and evaluation data for disfluency detection in learner speech. A comparative analysis of the NICT-JLE and Switchboard corpora confirmed the NICT-JLE corpus to be a suitable proxy for learner speech and identified multiple challenges for future disfluency detection models to address. The standardised train, heldout and test sets developed here not only allow for a fair comparison between any future models that are developed but also provide a framework for expanding model evaluation to areas that are important for language learning applications.

Limitations

The main limitations of this work are due to the original labelling approach taken for the development of the NICT-JLE corpus, the first of which is the labelling of learner errors. With only a subset of files labelled for errors, together with errors that occur in the reparandum going unlabelled, there is a limited scope for model development or analysis using these features. As explored above, errors can occur as part of disfluencies, and having a full dataset labelled with these occurrences would be valuable, not only for evaluation but also for joint modelling of the linguistic phenomena. The second limitation is that the approach to disfluency labelling is not directly comparable to that of the Switchboard corpus, as the NICT-JLE corpus does not contain labels for repair phrases and follows different nesting rules than those set out by Shriberg (1994).

The NICT-JLE corpus is additionally limited in that it only contains speech from Japanese-speaking learners of English. With the knowledge that learners' native language can influence factors such as disfluency frequency (Zuniga and Simard, 2019), it would be valuable to collect data from learners with varied first language backgrounds. In a similar vein, using a POS tagger specifically developed for learner speech such as that described by Nagata et al. (2018) would be beneficial for future iterations of the adapted corpus.

Acknowledgements

Funding for this work was awarded through the University of Sheffield Publication Scholarship scheme.

References

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

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

Andrew Caines, Diane Nicholls, and Paula Buttery. 2017. Annotating errors and disfluencies in transcriptions of speech. Technical Report UCAM-CL-TR-915, University of Cambridge, Computer Laboratory.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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

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

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.

Rie Koizumi and Yo In'nami. 2013. Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching & Research*, 4(5).

Judit Kormos. 1998. A new psycholinguistic taxonomy of self-repairs in l2: A qualitative analysis with retrospection. *Even Yearbook, ELTE SEAS Working Papers in Linguistics*, 3:43–68.

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

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.

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. 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.

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

Lucy Skidmore and Roger Moore. 2022. Incremental disfluency detection for spoken learner English. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 272–278, Seattle, Washington. Association for Computational Linguistics.

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.

Gerdina Wilhelmina Carla Maria Van Hest. 1996. *Self-repair in L1 and L2 production*. Tilburg: Tilburg University Press.

Richard Wiese. *Language production in foreign and native languages: Same or different*, pages 11–25. Narr Verlag.

Michael Zuniga and Daphnée Simard. 2019. Factors influencing l2 self-repair behavior: The role of l2 proficiency, attentional control and l1 self-repair behavior. *Journal of psycholinguistic research*, 48(1):43–59.