The effect of disfluencies and learner errors on the parsing of spoken learner language

Andrew Caines **Paula Buttery**
Institute for Automated Language Teaching and Assessment
Department of Theoretical and Applied Linguistics
University of Cambridge, Cambridge, U.K.
`(apc38|pjb48)@cam.ac.uk`

Abstract

NLP tools are typically trained on *written* data from *native speakers*. However, research into language acquisition and tools for language teaching & proficiency assessment would benefit from accurate processing of *spoken* data from *second language learners*. In this paper we discuss manual annotation schemes for various features of spoken language; we also evaluate the automatic tagging of one particular feature (filled pauses) – finding a success rate of 81%; and we evaluate the effect of using our manual annotations to 'clean up' the transcriptions for sentence parsing, resulting in a 25% improvement in parse success rate by completely cleaning the texts of disfluencies and errors. We discuss the need to adapt existing NLP technology to non-canonical domains such as spoken learner language, while emphasising the worth of continued integration of manual and automatic annotation.

1 Introduction

Natural language processing (NLP) tools are typically trained on *written* data from *native speakers*. However, research into language acquisition and tools for language proficiency assessment & language teaching – such as learner dialogue and feedback systems – would benefit from accurate processing of *spoken* data from *second language learners*. Being able to convert the text from unparseable to parseable form will enable us to (a) posit a target hypothesis that the learner intended to produce, and (b) provide feedback on this target based on the information removed or repaired in achieving that parseable form.

To proceed towards this goal, we need to adapt current NLP tools to the non-canonical domain of spoken learner language in a persistent fashion rather than use ad hoc post-processing steps to 'correct' the non-canonical data. Outcomes of this approach have been reported in the literature (*e.g.* Rimell & Clark (2009) in the biomedical domain; Caines & Buttery (2010) for spoken language). These fully adaptive approaches require large amounts of annotated data to be successful and, as we intend to work along these lines in future, the discussion in this paper is pointed in that direction.

The work presented here will act as a foundation for more permanent adaptations to existing tools. We annotate transcriptions of speech for linguistic features that are known to interfere with standard NLP to assess whether large-scale annotation of these features will be useful for training purposes. Obvious instances of this include disfluencies (*e.g.* filled pauses, false starts, repetition), formal errors of morphology and syntax, as well as 'errors' of word and phrase selection[1].

Since manual annotation is costly in terms of time and often money, one might question whether so many feature types are strictly necessary or even helpful for the task in hand. Indeed, filled pauses such as 'oh' and 'um' are already accounted for in the part-of-speech (POS) tagset we use (CLAWS2 (Garside, 1987)); and one might also argue that lexico-semantic errors might be dismissed *a priori* on the assumption that both the original and proposed forms are of the same POS (and thus won't affect a parser that performs tagging before the parse). We investigate the contribution of these features to

[1] The word 'error' appears here in quotes as it might be argued that questionable lexical selections are more a matter of infelicity and improbability than any strict dichotomy; we put this concern aside for now as a matter for future research.

First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages, pages 74–81,
Dublin, Ireland, August 23-29 2014.

parsing success. From a theoretical perspective we are interested in these features with regard to second language acquisition and therefore need to analyse them closely.

In this paper we describe our initial efforts to address the challenge of parsing learner speech with tools trained on native speaker writing. We also present empirical results that demonstrate the utility of annotated spoken transcription with respect to both tagging and parsing. We investigate: [i] the frequency of disfluencies, formal errors of morpho-syntax, and idiomatic errors of lexico-semantics in a corpus of spoken learner language; [ii] the accuracy of part-of-speech labels produced by the tagger associated with the Robust Accurate Statistical Parsing System (RASP (Briscoe et al., 2006)) for a particular type of disfluency (the filled pause)[2]; [iii] parse success rates and parse likelihoods using the RASP System with the texts in various 'modes' ranging from unaltered transcription to fully edited and corrected[3].

We find that in our spoken learner corpus of 2262 words, (i) around a quarter of words are annotated as disfluencies or errors; (ii) 81% of filled pauses were correctly tagged, meaning 1 in 5 are incorrectly tagged; (iii) mean parse likelihood for the text 'as is', unaltered, is –2.599 with a parse success rate of 47%, whereas completely 'cleaned up' text improves those scores to –1.995 and 72%[4]. We discuss the implications of these results below, along with the background context for our study and a more detailed description of our investigations.

2 Background

Previous analyses of the NLP of learner language include various experiments on the tagging and parsing of errorful text. Geertzen et al. (2013) employed the Stanford parser on written English learner data and achieved labelled and unlabelled attachment scores (LAS, UAS)[5] of 89.6% and 92.1%. They found that errors at the morphological level lead to incorrect POS-tagging, which in turn can result in an erroneous parse. Others have focused only on the POS-tagging of written learner corpora – for example with English (van Rooy and Schäfer, 2003) and French learner data (Thouësny, 2011) – demonstrating that post-hoc corrections for the most frequent tagging errors results in significant parse error reduction.

In other investigations of standard NLP tools on learner corpora, Ott & Ziai (2010) report general robustness using MaltParser on written German learner language; however, they found that by manually correcting POS tags, LAS improved from 79.15% to 85.71% and UAS from 84.81% to 90.22%. Wagner & Foster (2009) ran a series of parsing experiments using parallel errorful/corrected corpora, including a spoken learner corpus in which the likelihood of the highest ranked tree for corrected sentences was higher than that of uncorrected sentences in 69.6% of 500 instances. Taken together, these studies suggest that existing NLP tools remain robust to learner data, even more so if the original texts can be corrected and if the tagging stage is in some way verified, or adapted (*e.g.* Zinsmeister et al. (2014)).

On the other hand, Díaz-Negrillo et al. (2010) argue that treating learner data as a 'noisy variant' of native language glosses over systematic differences between the two, and instead call for a dedicated tagging format for 'interlanguage', one that encodes distributional, morphological and lexical information. For instance, 'walks' in 'John has walks' would morphologically be tagged as a present tense 3rd-person verb, but distributionally tagged as a past participle[6]. This is the kind of adaptation of existing tools that we advocate, though we would add that this system should be available for not just interlanguage but all data, allowing for non-canonical language use by native speakers as much as learners.

As for spoken language, Caines & Buttery (2010) among others suggest that adaptation can also be made to the parser, such that it enters a 'speech-aware mode' in which the parser refers to additional and/or replacement rules adapted to the particular features of spoken language. They demonstrated this with the omission of auxiliary verbs in progressive aspect sentences ('you talking to me?', 'how you

[2] The RASP POS-tagger was evaluated on the 560 randomly selected sentences from *The Wall Street Journal* that constitute the PARC dependency bank (DepBank; (King et al., 2003)) and achieved 97% accuracy (Briscoe et al., 2006).

[3] The RASP parser achieves a 79.7% microaveraged F_1 score on grammatical relations in DepBank (Briscoe and Carroll, 2006).

[4] *N.B.* the closer the parse likelihood to zero, the more probable the parse in the English language.

[5] LAS indicates the proportion of tokens that are assigned both the correct head and the correct dependency label; UAS indicates the proportion of tokens assigned the correct head, irrespective of dependency label.

[6] We thank reviewer #1 for this example.

doing?') and achieved a 30% improvement in parsing success rate for this construction type.

3 The corpus

Our speech data consist of recordings from Business Language Testing Service (BULATS) speaking tests[7]. In the test, learners are required to undertake five tasks; we exclude the tasks involving brief question-answering ('can you tell me your full name?', 'where are you from?', etc) and elicited imitation, leaving us with three free-form speech tasks. For this particular test the tasks were: [a] talk about some advice from a colleague (monologue), [b] talk about a series of charts from *Business Today* magazine (monologue), [c] give advice on starting a new retail business (dialogue with examiner).

In our full dataset the candidates come from India, Pakistan and Brazil, with various first languages (L1) including Hindi, Gujarati, Malayalam, Urdu, Pashto and Portuguese, and an age range of 16 to 47 at the time of taking the test. However, in this analysis we have only sampled recordings from candidates deemed to be at 'B2' upper intermediate level on the CEFR scale[8], so that the proficiency level of language used (and how that relates to NLP) is controlled for. In addition the L1s in our sample are Gujarati, Punjabi and Urdu only. This gives us a sample corpus of 2262 tokens in 'as-is' format (*i.e.* the true transcriptions before any corrections are made).

4 Manual annotation

The recordings were manually transcribed and annotated for various features falling into three categories described and exemplified in the following non-exhaustive list.

- *disfluencies* – interruptions to the flow of otherwise fluent speech;
  - <fp> filled pauses (tokens such as *uh, er, um* that serve to fill time and hold the turn) and <rep n="n"> repetition (the speaker repeats a word or phrase one or several times):
    "or the other way is to <fp>um</fp> <rep n="1">is to</rep> raise finance"
  - <false> false starts – the speaker begins to express a word or phrase which he then corrects:
    "in two thousand eight it was <false>thirty five p</false> thirty percent"
- *formal errors* of morpho-syntax, such as number agreement, verb inflection and word order errors;
  - noun form: "for becoming a chartered <NS type="FN"><i>accountants</i><c>accountant </c></NS>"
  - missing verb: "as the charts <NS type="MV"><c>show</c></NS> its sales increased"
  - word order: "<NS type="W"><i>how it would be help for you mention</i><c>mention how it helped you</c></NS>"
- *idiomatic 'errors'* – infelicities in lexical selection, failure to express intended meaning, or less-than-natural phrasing;
  - idiomatic: "all my class <NS type="ID"><i>fellows</i><c>mates</c></NS>"
  - idiomatic: "to <NS type="ID"><i>get in</i><c>make a</c></NS> profit"
  - replace quantifier: "for a bank to grant us <NS type="RQ"><i>some</i><c>a</c></NS> loan"

The annotation scheme for formal and idiomatic errors comes from the project to annotate the Cambridge Learner Corpus (Briscoe et al., 2010). The 'error zone' is denoted by <NS> tags, with any original token(s) enclosed by <i> and any proposed correction enclosed by <c>. The various error types are defined in Nicholls (2003) and the categories are similar to the ones given: either self-defining ('ID' for idiom error, 'W' for word order, *etc*) or a combination of operation plus part-of-speech ('FN' form of noun, 'MV' missing verb, 'RQ' replace quantifier, *etc*).

In Table 1 we report the number of errors and disfluencies found in our corpus along with a relative frequency per 100 words. Just under a quarter of the thousand tokens in our corpus are affected by disfluencies and errors, with the former being far more prevalent.

[7] We thank Cambridge English Language Assessment for releasing these recordings for this pilot study; for further information on BULATS go to `http://www.bulats.org/`

[8] The 'Common European Framework of Reference for Languages': a schema for grading an individual learner's language level. For further information go to `http://www.coe.int/lang-CEFR`

All transcription and annotation has been carried out by a single annotator (the first author). It would be interesting to obtain measures of inter-annotator agreement to assess the extent to which the nature of error judgement (particularly in judgements as to idiomaticity) is subjective.

type	**instances in corpus**	**relative frequency (per 100 words)**
disfluency	316	14
formal error	143	6
idiomatic error	70	3
total	529	23

Table 1: Error counts in our corpus

5 Automated annotation: part-of-speech tagging

Since filled pauses such as 'er' and 'um' are included in the CLAWS2 tagset used by the RASP System as UH, 'interjection', one might question the worth of manually annotating filled pauses (FPs). Of the disfluency set, it might be one small time-saving to leave these to the tagger. However, 'interjection' is not a homogeneous set, as UH also covers exclamations of surprise ('oh') and assent ('yes'). Moreover, we find that the POS-tagging of tokens annotated as FPs is not entirely appropriate in this non-canonical domain. Table 2 shows that the majority of FPs are correctly tagged UH, though others are tagged as nouns (NN), verbs (VV), adjectives (JJ), adverbs (RR) and foreign words (&FW)[9].

Token	UH	&FW	JJ	NN	RR	VV	*total*
er	104	0	0	0	0	0	104
mm	0	0	0	8	0	0	8
uh	0	0	0	1	2	4	7
um	2	5	0	0	0	0	7
nuh	0	0	0	0	2	1	3
buh	0	0	0	1	0	1	2
other	2	0	1	0	0	0	3
total	108	5	1	10	6	9	134

Table 2: POS tagging of filled pauses

One possible solution is to append a dictionary of known FP tokens to the tagger, and specify that they should be tagged UH, or even better, a new tag such as FP. But as the Table demonstrates, there are standard, highly frequent FPs such as 'er', 'uh' and 'um', and then there are novel forms such as 'nuh', 'buh' and 'nna' which we found to be rather idiosyncratic – *i.e.* there might be novel FPs for every individual. Moreover, the introduction of a closed class depends on consistent transcription practice, not necessarily a given with even a lone annotator, let alone more than one.

Automatic identification and repair of disfluencies is a well-developed research topic, with continuing refinements to joint parsing and disfluency detection models (*e.g.* Qian & Liu (2013), Rasooli & Tetreault (2014), Honnibal & Johnson (2014)), plus applied work in the domains of automatic speech recognition (Fitzgerald et al., 2009) and machine translation (Cho et al., 2014). We note the linguistic rules included in the Lease, Johnson & Charniak (2006) tree adjoining grammar (TAG) noisy-channel model – lexical, POS and syntactic rules that reduce errors in the TAG model. This is another case of improvements to NLP tools thanks to data-driven linguistic insight, and a design that we could incorporate into our work on automated assessment and feedback.

[9] The 'other' filled pauses are singleton forms: *eh, nna, ah*.

6 Automated annotation: sentence parsing

In this section we report the results of our parsing experiment in which transcribed learner utterances were processed by the RASP system in four different forms:

(A) as-is: without alteration;

(B) less-disfluency: with disfluencies removed;

(C) less-form-error: with morpho-syntactic errors corrected;

(D) less-lex-error: with semantic/idiomatic improvements.

We investigated the effect on the parsing output of each transcription format compared to the (A) format as a baseline. We processed each format in turn singularly, as well as cumulative combinations of (B), (C) and (D) in every possible order. The results are set out in Table 3, with mean likelihoods of the highest ranked parse for each sentence (μ)[10], differences between this mean and the baseline where applicable (Δ_{base}), and success rates in terms of non-fragmentary tree outputs (*i.e.* parses labelled other than 'T/frag' in the RASP System).

mode	μ	Δ_{base}	$\neg$T/frag	**mode**	μ	Δ_{base}	$\neg$T/frag	**mode**	μ	Δ_{base}	$\neg$T/frag
(A)	–2.599	0	.471	(A)	–2.599	0	.471	(A)	–2.599	0	.471
(B)	–2.094	+.505	.623	(BC)	–2.032	+.567	.689	(BCD)	–1.995	+.604	.715
(C)	–2.574	+.025	.484	(BD)	–2.049	+.550	.649		-	-	-
(D)	–2.563	+.036	.503	(CD)	–2.545	+.054	.523		-	-	-

Table 3: Mean parse likelihoods, deltas to baseline and parse success rates in all transcription modes

As can be seen in Table 3, the removal of disfluencies (B) is the single move of greatest benefit to parse likelihood scores and parse tree success rates compared to the 'as-is' baseline (A). The correction of morpho-syntactic (C) and idiomatic errors (D) have a lesser effect. All pairings have a positive effect on parse likelihoods, especially those featuring disfluency removal (B); and the three 'corrective' steps combined (BCD) have the greatest effect of all.

However, we show by analysis of two candidates in our corpus that these effects can differ on an individual basis. In Figure 1, the candidate on the left has a less pronounced effect of disfluency removal (B) compared to the baseline (A) than the candidate on the right. The effect of both formal (C) and idiomatic (D) error correction are also seen to make improvements over (A), which is not the case for the second candidate. Such observations serve as a reminder that when generalising about overall corpus patterns we collapse over so many individual language models. It may well turn out that disfluencies are an especially idiosyncratic type of language use, an avenue we will explore in future work.

7 Discussion

In this paper we have investigated NLP of transcribed learner speech, questioning how tools trained on native speaker written data would handle such data. We found that the majority (81%) of filled pauses were correctly tagged 'UH', though this only covers three of eleven FP forms (*er, um, eh*). We propose a dictionary of FPs and a specific FP POS-tag, while suggesting that the dictionary will not catch all novel FPs (since they seem to be idiosyncratic) and that we can turn to state-of-the-art research on automated disfluency detection to help us.

We also showed that sentence parsing could be improved from a 47% 'success' rate (*i.e.* non-fragmentary (T/frag in RASP parlance) parse trees) in the 'as-is' transcriptions, to 72% in transcriptions with disfluencies removed and errors corrected (see Table 3). We found that disfluency removal is the main contributor to this improvement, though this was found to be somewhat idiosyncratic (as in Figure 1).

[10] Note that parse likelihoods have been normalised for word length, as they increase in a near-linear manner according to the number of terminal nodes in a tree.

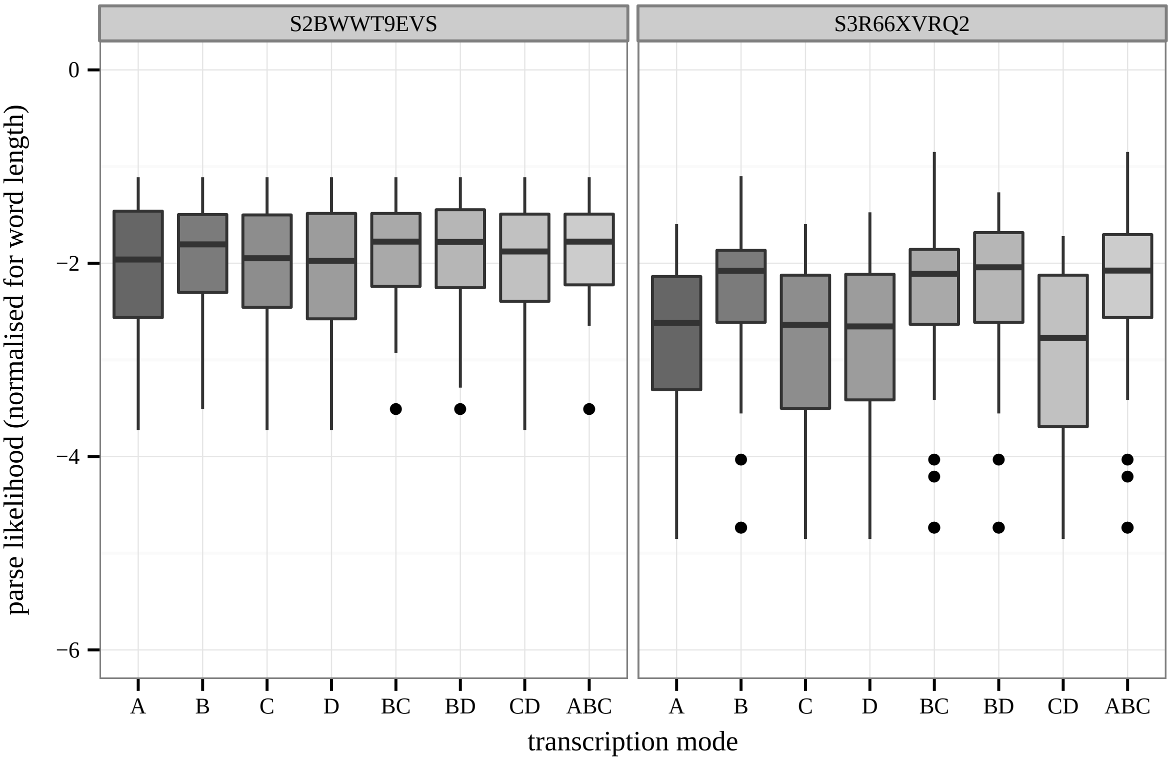

Figure 1: Parse likelihoods for each transcription mode, for two individuals in our corpus; the whiskers indicate the largest and smallest observation within $1.5 * IQR$ (inter-quartile range; the distance between first and third quartiles), while the upper hinge indicates the third quartile (75th percentile), the middle is the median, the lower hinge is the first quartile (25th percentile), and the points are outliers.

The motivation for this work is to investigate what is required to convert texts from unparseable to parseable form. The steps taken to achieve this can be used to inform automated learner dialogue or feedback systems. We note that automated assessment may be improved by parse trees but may well be performed without them: it can proceed on the basis of superficial detection of features known to correlate with high grades (possibly including certain disfluency types, for instance). But to be able to diagnose *how the learner can improve*, we need a deeper structural analysis of the text – *i.e.* requiring that the text is in parseable form. Our manual annotations are one step towards this goal.

Our annotations also indicate that spoken learner data features many disfluencies and errors, with over a quarter of the 2262-word testset affected in some way. Automatic error detection (and correction) is a burgeoning field (see for example the work on learner data by Briscoe et al. (2010), Andersen (2011) and Kochmar & Briscoe (2014), as well as the most recent shared task on grammatical error correction at CoNLL-2014 (Ng et al., 2014)). Such studies are based on written language. We envisage adding speech-specific information and adaptations to such systems on the basis of our fuller annotation project.

Indeed, it so happens that the problem of NLP in the spoken domain is one we address here with learner data. However, we do not assume that the problem of adapting or building NLP tools for spoken data is substantially different for native speaker data. We intend to collect recordings of native speakers undertaking the same tasks as the BULATS candidates, allowing for comparative studies of errors and disfluencies in native and learner data, with the task and topic variables held constant as far as possible.

Finally, we emphasise that we intend to add to the corpus with more annotated data from a wider range of L1s and a wider range of proficiency levels. We can then investigate the possible effects of more varied syntactic complexity, lexical diversity and error types.

Acknowledgments

We thank Cambridge English Language Assessment for funding this work and providing the data. We also thank Ted Briscoe, Mike McCarthy and Øistein Andersen for their support and advice, and we are grateful to the three anonymous reviewers for their helpful comments and suggestions.

References

Øistein E. Andersen. 2011. Semi-automatic ESOL error annotation. *English Profile Journal*, 2:e1.

Ted Briscoe and John Carroll. 2006. Evaluating the accuracy of an unlexicalized statistical parser on the PARC DepBank. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Association for Computational Linguistics.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP System. In *Proceedings of the COLING/ACL 2006 Interactive Presentations Session*. Association for Computational Linguistics.

Ted Briscoe, Ben Medlock, and Øistein E. Andersen. 2010. Automated assessment of ESOL free text examinations. *University of Cambridge Computer Laboratory Technical Reports*, 790.

Andrew Caines and Paula J. Buttery. 2010. 'You talking to me?' A predictive model for zero auxiliary constructions. In *Proceedings of the Workshop on Natural Language Processing and Linguistics, Finding the Common Ground, Annual Meeting of the Association for Computational Linguistics (ACL) 2010*. Association for Computational Linguistics.

Eunah Cho, Jan Niehues, and Alex Waibel. 2014. Tight integration of speech disfluency removal into SMT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*. Association for Computational Linguistics.

Ana Díaz-Negrillo, Detmar Meurers, Salvador Valera, and Holger Wunsch. 2010. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36:139–154.

Alison Edwards. 2014. The progressive aspect in the Netherlands and the ESL/EFL continuum. *World Englishes*, 33:173–194.

Erin Fitzgerald, Keith Hall, and Frederick Jelinek. 2009. Reconstructing false start errors in spontaneous speech text. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*. Association for Computational Linguistics.

Roger Garside. 1987. The CLAWS word-tagging system. In Roger Garside, Geoffrey Leech, and Geoffrey Sampson, editors, *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic linguistic annotation of large scale L2 databases: the EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum*. Somerville, MA: Cascadilla Proceedings Project.

Matthew Honnibal and Mark Johnson. 2014. Joint incremental disfluency detection and dependency parsing. *Transactions of the Association for Computational Linguistics*, 2:131–142.

Tracy H. King, Richard Crouch, Stefan Riezler, Mary Dalrymple, and Ronald M. Kaplan. 2003. The PARC700 Dependency Bank. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC 2003)*.

Ekaterina Kochmar and Ted Briscoe. 2014. Detecting learner errors in the choice of content words using compositional distributional semantics. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*. Association for Computational Linguistics.

Matthew Lease, Mark Johnson, and Eugene Charniak. 2006. Recognizing disfluencies in conversational speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:1566–1573.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Eighteenth Conference on Computational Natural Language Learning, Proceedings of the Shared Task*. Association for Computational Linguistics.

Diane Nicholls. 2003. The Cambridge Learner Corpus: error coding and analysis for lexicography and ELT. In Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery, editors, *Proceedings of the Corpus Linguistics 2003 conference; UCREL technical paper number 16*. Lancaster University.

Niels Ott and Ramon Ziai. 2010. Evaluating dependency parsing performance on German learner language. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (NEALT 2010)*.

Xian Qian and Yang Liu. 2013. Disfluency detection using multi-step stacked learning. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Mohammad Sadegh Rasooli and Joel Tetreault. 2014. Non-monotonic parsing of *Fluent umm I Mean* disfluent sentences. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*. Association for Computational Linguistics.

Laura Rimell and Stephen Clark. 2009. Porting a lexicalized-grammar parser to the biomedical domain. *Journal of Biomedical Informatics*, 42:852–865.

Sylvie Thouësny. 2011. Increasing the reliability of a part-of-speech tagging tool for use with learner language. In *Proceedings of the Pre-conference Workshop on Automatic Analysis of Learner Language, CALICO Conference 2009*.

Bertus van Rooy and Lande Schäfer. 2003. An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus. In *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University.

Joachim Wagner and Jennifer Foster. 2009. The effect of correcting grammatical errors on parse probabilities. In *Proceedings of the 11th International Conference on Parsing Technologies*.

Heike Zinsmeister, Ulrich Heid, and Kathrin Beck. 2014. Adapting a part-of-speech tagset to non-standard text: the case of STTS. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. European Language Resources Association.