Parsing transcripts of speech

Andrew Caines
ALTA Institute
University of Cambridge
apc38@cam.ac.uk

Michael McCarthy
School of English
University of Nottingham
mactoft@cantab.net

Paula Buttery
Computer Laboratory
University of Cambridge
pjb48@cam.ac.uk

Abstract

We present an analysis of parser performance on speech data, comparing word type and token frequency distributions with written data, and evaluating parse accuracy by length of input string. We find that parser performance tends to deteriorate with increasing length of string, more so for spoken than for written texts. We train an alternative parsing model with added speech data and demonstrate improvements in accuracy on speech-units, with no deterioration in performance on written text.

1 Introduction

Relatively little attention has been paid to parsing spoken language compared to parsing written language. The majority of parsers are built using newswire training data and *The Wall Street Journal* section 21 of the Penn Treebank is a ubiquitous test set. However, the parsing of speech is of no little importance, since it's the primary mode of communication worldwide, and human computer interaction through the spoken modality is increasingly common.

In this paper we first describe the morpho-syntactic characteristics of spoken language and point out some key distributional differences with written language, and the implications for parsing. We then investigate how well a commonly-used open source parser performs on a corpus of spoken language and corpora of written language, showing that performance deteriorates sooner for speech as the length of input string increases. We demonstrate that a new parsing model trained on both written and spoken data brings improved performance, making this model freely available[1]. Finally we consider a modification to deal with long input strings in spoken language, a preprocessing step which we plan to implement in future work.

[1] https://goo.gl/iQMu9w

2 Spoken language

As has been well described, speech is very different in nature to written language (Brazil, 1995; Biber et al., 1999; Leech, 2000; Carter and McCarthy, 2017). Putting aside the mode of transmission for now – the phonetics and prosody of producing speech versus the graphemics and orthography of writing systems – we focus on morphology, syntax and vocabulary: that is, the components of speech we can straightforwardly analyse in transcriptions. We also put aside pragmatics and discourse analysis therefore, even though there is much that is distinctive in speech, including intonation and co-speech gestures to convey meaning, and turn-taking, overlap and co-construction in dialogic interaction.

A fundamental morpho-syntactic characteristic of speech is the lack of the sentence unit used by convention in writing, delimited by a capital letter and full stop (period). Indeed it has been said that, "such a unit does not realistically exist in conversation" (Biber et al., 1999). Instead in spoken language we refer to 'speech-units' (SUs)– token sequences which are usually coherent units from the point of view of syntax, semantics, prosody, or some combination of the three (Strassel, 2003). Thus we are able to model SU boundaries probabilistically, and find that, in dialogue at least, they often coincide with turn-taking boundaries (Shriberg et al., 2000; Lee and Glass, 2012; Moore et al., 2016).

Other well-known characteristics of speech are disfluencies such as hesitations, repetitions and false starts (1)-(3).

(1) um he's a closet yuppie is what he is (Leech,

2000).

(2) I played, I played against um (Leech, 2000).

(3) You're happy to – welcome to include it (Levelt, 1989).

Disfluencies are pervasive in speech: of an annotated 767k token subset of the Switchboard Corpus of telephone conversations (SWB), 17% are disfluent tokens of some kind (Meteer et al., 1995). Furthermore they are known to cause problems in natural language processing, as they must be incorporated in the parse tree or somehow removed (Nasr et al., 2014). Indeed an 'edit' transition has been proposed specifically to deal with automatically identified disfluencies, by removing them from the parse tree constructed up to that point along with any associated grammatical relations (Honnibal and Johnson, 2014; Moore et al., 2015).

We compared the SWB portion of Penn Treebank 3 (Marcus et al., 1999) with the three English corpora contained in Universal Dependencies 2.0 (Nivre et al., 2017) as a representation of the written language. These are namely:

- The 'Universal Dependencies English Web Treebank' (EWT), the English Web Treebank in dependency format (Bies et al., 2012; Silveira et al., 2014);
- 'English LinES' (LinES), the English section of a parallel corpus of English novels and Swedish translations (Ahrenberg, 2015);
- The 'Treebank of Learner English' (TLE), a manually annotated subset of the Cambridge Learner Corpus First Certificate in English dataset (Yannakoudakis et al., 2011; Berzak et al., 2016).

We found several differences between our spoken and written datasets in terms of morphological, syntactic and lexical features. Firstly, the most frequent tokens in writing (ignoring punctuation marks) are, unsurprisingly, function words – determiners, prepositions, conjunctions, pronouns, auxiliary and copula verbs, and the like (Table 1). These are normally considered 'stopwords' in large-scale linguistic analyses, but even if they are semantically uninteresting, their ranking is indicative of differences between speech and writing.

Speech	Freq.	Rank	Writing	Freq.
I	46,382	1	the	41,423
and	33,080	2	to	26,459
the	29,870	3	and	22,977
you	27,142	4	I	20,048
that	27,038	5	a	18,289
it	26,600	6	of	18,112
to	22,666	7	in	14,490
a	22,513	8	is	10,020
uh	20,695	9	you	10,002
's	20,494	10	that	9952
of	17,112	11	for	8578
yeah	14,805	12	it	8238
know	14,723	13	was	8195
they	13,147	14	have	6604
in	12,548	15	on	5821
do	12,507	16	with	5621
n't	11,100	17	be	5514
we	10,308	18	are	4815
have	9970	19	not	4716
uh-huh	9325	20	my	4478

Table 1: The most frequently occurring tokens in selected corpora of English speech (the Switchboard Corpus in Penn Treebank 3) and writing (EWT, LinES, TLE), normalised to counts per million.

In SWB the most frequent token is *I* followed by *and*, then *the* albeit much less frequently than in writing, then *you, that, it* at much higher relative frequencies (per million tokens) than in writing. This ranking reflects the way that (telephone) conversations revolve around the first and second person (*I* and *you*), and the way that speech makes use of coordination and hence the conjunction *and* much more than writing.

Furthermore clitics indicative of possession, copula or auxiliary *be*, or negation (*'s, n't*) and discourse markers *uh, yeah, uh-huh* are all in the twenty-five most frequent terms in SWB. The single content word in these top-ranked tokens (assuming *have* occurs mainly as an auxiliary) is *know*, 13th most frequent in SWB, but as will become clear in Table 3, it's hugely boosted by its use in the fixed phrase, *you know*.

Finally we note that the normalised frequencies for these most frequent tokens are higher in speech than in writing, suggesting that there is greater distributional mass in fewer token types in SWB, a suggestion borne out by sampling 394,611 tokens (the sum total of the three written corpora) from SWB 100 times and finding that not once does the vocabulary size exceed even half that of the written corpora (Table 2).

With the most frequent bigrams we note further differences between speech and writing (Ta-

Medium	Tokens	Types
speech	394,611*	11,326**
writing	394,611	27,126

Table 2: Vocabulary sizes in selected corpora of English speech and writing (* sampled from 766,560 tokens in SWB corpus; ** mean of 100 samples, st.dev=45.5).

ble 3). The most frequent bigrams in writing tend to be combinations of preposition and determiner, or pronoun and auxiliary verb. In speech on the other hand, the very frequent bigrams include the discourse markers *you know, I mean*, and *kind of*, pronoun plus auxiliary or copula *it's, that's, I'm, they're*, and *I've*, and disfluent repetition *I I*, and hesitation *and uh*. Again frequency counts are lower for the written corpus, symptomatic of a smaller set of bigrams in speech. There are 163,690 unique bigrams in the written data, and a mean of 89,787 (st.dev=151) unique bigrams in SWB from 100 samples.

Speech	Freq.	Rank	Writing	Freq.
you know	11,165	1	of the	4313
it's	8531	2	in the	3702
that's	6708	3	to the	2352
don't	5680	4	I have	1655
I do	4390	5	on the	1607
I think	4142	6	I am	1500
and I	3790	7	for the	1475
I'm	3716	8	I would	1427
I I	3000	9	and the	1389
in the	2972	10	and I	1361
and uh	2780	11	to be	1318
a lot	2714	12	I was	1140
of the	2655	13	don't	1125
it was	2616	14	will be	1092
I mean	2518	15	it was	1057
kind of	2448	16	at the	1044
they're	2349	17	in a	1041
I've	2165	18	like to	1036
going to	2135	19	is a	1021
lot of	2053	20	it is	998

Table 3: The most frequently occurring bigrams in selected corpora of English speech (the Switchboard Corpus in Penn Treebank 3) and writing (EWT, LinES, TLE), normalised to counts per million.

In Table 4 we present a short list of the most frequent dependency types, represented as part-of-speech tag pairs TAG_1_TAG_2, where TAG_1 is the head and TAG_2 is the dependent. In speech we see that several of the most frequent dependency pairs involve a verb or root as the head, whereas the most frequent pairs in writing involve a noun.

We are certain that in future work there are further insights to be gleaned from comparisons of speech and writing at higher-order *n*-grams and in terms of dependency relations between tokens. These may in turn have implications for parsing algorithms, or at least may suggest some solutions for more accurate parsing of speech. Other genres and styles of speech and writing would also be worthy of study – especially more recently collected recordings of speech.

Speech	Freq.	Rank	Writing	Freq.
VBP_PRP	51,845	1	NN_DT	48,846
NN_DT	47,469	2	NN_IN	36,274
ROOT_UH	39,067	3	NN_NN	27,490
IN_NN	26,868	4	NN_JJ	21,566
VB_PRP	24,321	5	VB_NN	19,584
ROOT_VBP	24,156	6	VB_PRP	16,320

Table 4: The most frequently occurring part-of-speech tag dependency pairs in selected corpora of English speech (the Switchboard Corpus in Penn Treebank 3) and writing (EWT, LinES, TLE), normalised to counts per million. The first tag in the pair is the head of the relation; the second is the dependent (Penn Treebank tagset).

3 Parsing experiments

We used the Stanford CoreNLP toolkit (Manning et al., 2014) to tokenize, tag and parse input strings from a range of corpora. This includes the 766k token section of the Switchboard Corpus of telephone conversations (SWB) distributed as part of Penn Treebank 3 (Godfrey et al., 1992; Marcus et al., 1999), and English treebanks from the Universal Dependencies release 2 (Nivre et al., 2017). All treebanks are in CoNLL format[2] and we measure performance through unlabelled attachment scores (UAS) which indicate the proportion of tokens with correctly identified heads in the output of the parser, compared with gold-standard annotations (Kübler et al., 2009).

In Table 5 we report UAS scores overall for each corpus, along with corpus sizes in terms of tokens and sentence or speech units. It is apparent that (a) parser performance for speech units is much poorer than for written units, and that (b) performance across written corpora is broadly similar, though TLE (surprisingly) has the highest UAS score – possibly reflective of a tendency for language learners to write in syntactically more con-

[2]We thank Matthew Honnibal for sharing the SWB treebank converted to CoNLL-X format, arising from his TACL paper with Mark Johnson (Honnibal and Johnson, 2014).

servative ways [an issue we won't explore further here].

Corpus	Medium	Units	Tokens	UAS
SWB	speech	102,900	766,560	.540
EWT	writing	14,545	218,159	.744
LinES	writing	3650	64,188	.758
TLE	writing	5124	96,180	.845

Table 5: Corpus sizes and overall unlabelled attachment scores using Stanford Core NLP; SWB=Switchboard, EWT=English Web Treebank, LinES=English section LinES, TLE=Treebank of Learner English

Closer inspection of UAS scores by speech unit in SWB shows that parser performance is not uniformly worse than it is for written language. If we sort the input units into bins by unit length, we see that the parser is as accurate for shorter units of transcribed speech as it is for written units of similar lengths (Table 6)[3]. Indeed for speech units of 1-10 tokens in SWB, mean UAS is similar to that for sentence units of 1-10 tokens in EWT. However, the main difference in UAS scores over increasingly long inputs is the rate of deterioration in parser performance: for speech units the drop-off in UAS scores is much steeper.

Even with strings up to 40 tokens in length, mean UAS remains within 10 points of that for the 1-10 token bin in the three written corpora. But for SWB, mean UAS by that point is less than 50%. In fact in the 11-20 token bin we already see a steep drop-off in parser performance compared to the shortest class of speech unit.

It is only above 50 tokens that EWT and LinES UAS means fall by more than 10 percentage points compared to the 1-10 token score; for TLE this is true above 60 tokens. By this stage we are dealing with small proportions of the written corpora: 96.9% of the units in EWT and 98.1% in LinES are of length 50 tokens or fewer, whilst 99.8% of units in TLE are 60 tokens or shorter (Figure 1).

For SWB the problem is more acute, with 25.5% of units at least 11 tokens long and scoring mean UAS 50% or less. Figure 2 illustrates the disparity with boxplots showing UAS medians (thick line), first and third quartiles ('hinges' at bottom and top of box), ±1.5 inter-quartile range from the hinge (whiskers), and outliers beyond this range. It is apparent that parser performance

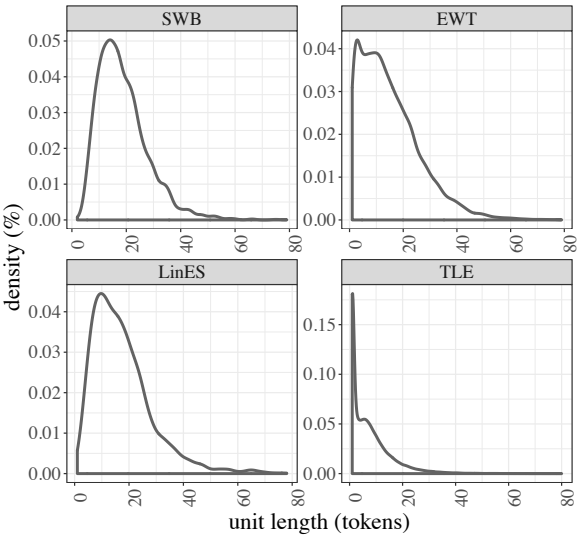

Figure 1: Density plot of unit lengths in four English corpora; SWB=Switchboard, EWT=English Web Treebank, LinES=English section LinES, TLE=Treebank of Learner English.

deteriorates as the unit length increases, for all corpora, but especially so for the speech corpus SWB.

What can be done to address this problem? One approach is to train a new parsing model on more appropriate training data, since general-purpose open-source parsers are usually trained on sections of *The Wall Street Journal* (WSJ) in Treebank 3 (Marcus et al., 1999). Training NLP tools with data appropriate to the medium, genre, or domain, is generally thought to be sensible and helpful to the task (Caines and Buttery, 2014; Plank, 2016). We do not claim this to be a groundbreaking proposal therefore, but instead present the results of such a step here for three reasons:

(i) To demonstrate how much improvement can be gained with a domain-appropriate parsing model;

(ii) To make the speech parsing model publicly available for other researchers;

(iii) To call for greater availability of speech transcript treebanks.

With regard to point (iii), to the best of our knowledge, the Switchboard portion of the Penn Treebank (PTB) is the only substantial, readily-available[4] treebank for spoken English. We welcome feedback to the contrary, and efforts to pro-

[3] Units longer than 80 tokens are omitted from the analysis as there are too few for meaningful comparison.

[4] Subject to licence available from the Linguistic Data Consortium.

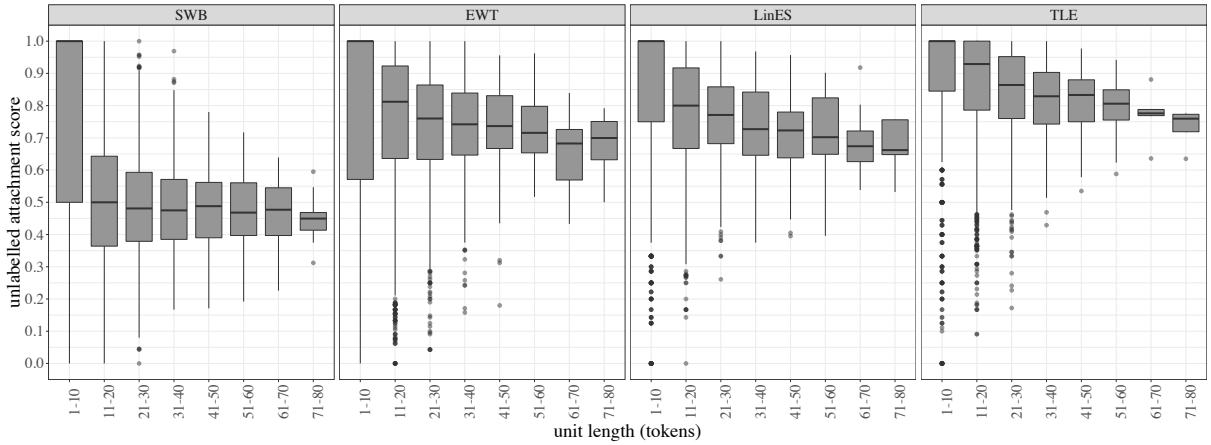

Figure 2: Unlabelled attachment scores by unit length in four English corpora.

	Unit length (tokens)							
Corpus	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80
SWB	.753 (76232)	.506 (19281)	.489 (4885)	.480 (1344)	.480 (366)	.473 (126)	.460 (37)	.447 (12)
EWT	.759 (6011)	.762 (4680)	.738 (2453)	.731 (944)	.736 (312)	.718 (96)	.655 (30)	.684 (12)
LinES	.826 (1086)	.770 (1433)	.761 (720)	.731 (251)	.710 (89)	.713 (37)	.674 (24)	.671 (5)
TLE	.866 (887)	.866 (2410)	.838 (1302)	.817 (380)	.816 (101)	.799 (34)	.770 (5)	.733 (4)

Table 6: Unlabelled attachment scores by unit length in four English corpora (number of units in parentheses).

duce new treebanks. Furthermore, if this is the situation for as well-resourced a language as English, we assume that the need for treebanks of speech corpora is even greater for other languages.

In point (ii) we don't imagine we're making a definitive statement on the best model for parsing speech – rather we think of it as a baseline against which future models can be compared. We welcome contributions in this respect.

As for point (i), we trained two new parsing models using the Stanford Parser (Klein and Manning, 2003). These were based on the WSJ sections of PTB as is standard, with added training data from SWB setting the maximum unit length first at 40 tokens – which appears to be the standard length for the models distributed with the parser – and secondly at an increased maximum of 80 tokens. Both were probabilistic context-free grammars. We refer to them as PCFG_WSJ_SWB_40 and PCFG_WSJ_SWB_80.

In Table 7 we show overall UAS scores for our four target English corpora, for three parsing models: the standard model distributed with CoreNLP, and our two new models, PCFG_WSJ_SWB_40 and PCFG_WSJ_SWB_80. It is apparent that the new models bring a large performance gain in parsing speech, as expected, plus a small performance gain in parsing writing – presumably because they can deal better than predominantly newswire trained models can with the less canonical syntactic structures contained in the written English obtained from the web and from learners. There is no apparent difference between PCFG_WSJ_SWB_40 and PCFG_WSJ_SWB_80 (therefore the latter does no harm and we make both available), presumably because there are relatively few units greater than 40 tokens and so any performance gain here has little bearing on the overall scores. Or, CoreNLP and PCFG_WSJ_SWB_40 are able to generalise to long strings as well as the PCFG_WSJ_SWB_80 model which has been presented with long string exemplars in training.

Model	**SWB**	**EWT**	**LinES**	**TLE**
CoreNLP	.540	.744	.758	.845
PCFG_WSJ_SWB_40	.624	.748	.760	.847
PCFG_WSJ_SWB_80	.624	.748	.760	.847

Table 7: Overall unlabelled attachment scores for four English corpora and three parsing models

In Figures 3 and 4 we show the difference between the CoreNLP and PCFG models in terms of UAS delta for each input unit. These are again binned by string length, and facetted by corpus. It is apparent that the alteration for the smallest units is somewhat volatile. This is understandable

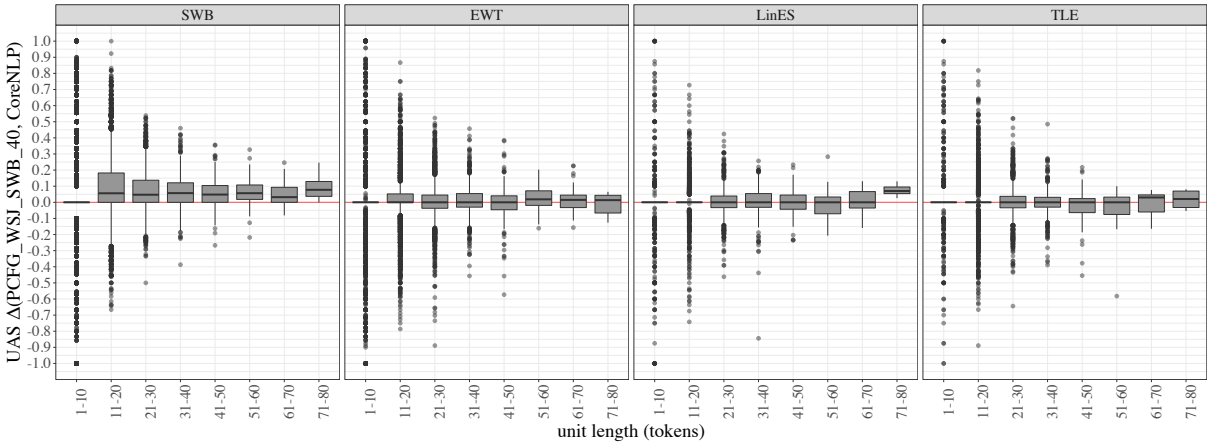

Figure 3: Unlabelled attachment scores by unit length in four English corpora: difference between model PCFG_WSJ_SWB_40 and CoreNLP.

given that a 1-token string which was correctly or incorrectly parsed by CoreNLP might now be incorrectly or correctly parsed by the PCFG models, leading to a delta of +1 or -1. Nevertheless the majority of short tokens are unaffected – shown by the median and hinges of the 1-10 token boxplot centring on y=0.

Where the added SWB training data seems to help is in units longer than 10 tokens, where the UAS delta median and hinges are consistently above zero, indicating improved performance. The boxplots tend to centre around zero for the written corpora, except for the 71-80 bin in LinES for which the boxplot is above zero, albeit for a small sample size of 5 (Table 6). The pattern for both PCFG models is broadly the same.

4 Related work

This is one among many studies examining the parsing of non-canonical data (Lease et al., 2006; Goldberg et al., 2014; Ragheb and Dickinson, 2014). Broadly speaking, there are two approaches to the problem (Eisenstein, 2013): (1) train new models specifically for non-canonical language; (2) normalise the data so that existing NLP tools work better on it. For example, Foster and colleagues (2008) deliberately introduced grammatical errors to copies of *WSJ* treebank sentences in order to train a parser to deal with noisy input. Daiber & van der Goot (2016), meanwhile, adopted the approach of text normalisation preceding syntactic parsing in dealing with social media data.

Some have proposed 'active learning' or 'self learning' algorithms for parser training, which learn from sparsely annotated or completely unannotated data (Mirroshandel and Nasr, 2011; Rei and Briscoe, 2013; Cahill et al., 2014). We could explore such methods for a speech-specific parser in future work, though they work better with large datasets to learn from – Rei & Briscoe trained on the 50 million token BLLIP corpus, for example. At the time of writing there are no similarly-sized speech corpora that we are aware of.

Relevant work on speech parsing includes that on automated disfluency detection and repair in speech transcriptions (Charniak and Johnson, 2001; Rasooli and Tetreault, 2013; Honnibal and Johnson, 2014; Moore et al., 2015; Yoshikawa et al., 2016), in which the problem has come to be addressed with a transition-based parser featuring an 'edit'-like action that can remove incrementally-constructed parse tree sections upon detection of a disfluency. Other approaches include prosodic information to detect disfluencies where the audio file is available alongside the transcription (Kahn et al., 2005). A combination of prosodic and morpho-syntactic features have been used to address another problem which affects parse quality: that of speech-unit delimitation, also known as 'speech segmentation' or 'sentence boundary detection' (Shriberg et al., 2000; Moore et al., 2016). SU delimitation and parsing were considered together as a joint problem, along with automatic speech recognition error rates, in a recent article by Kahn & Osterdorf (2012).

Finally, we should point out that we opted to work with Stanford CoreNLP for our parsing experiments because it is well-documented and well-

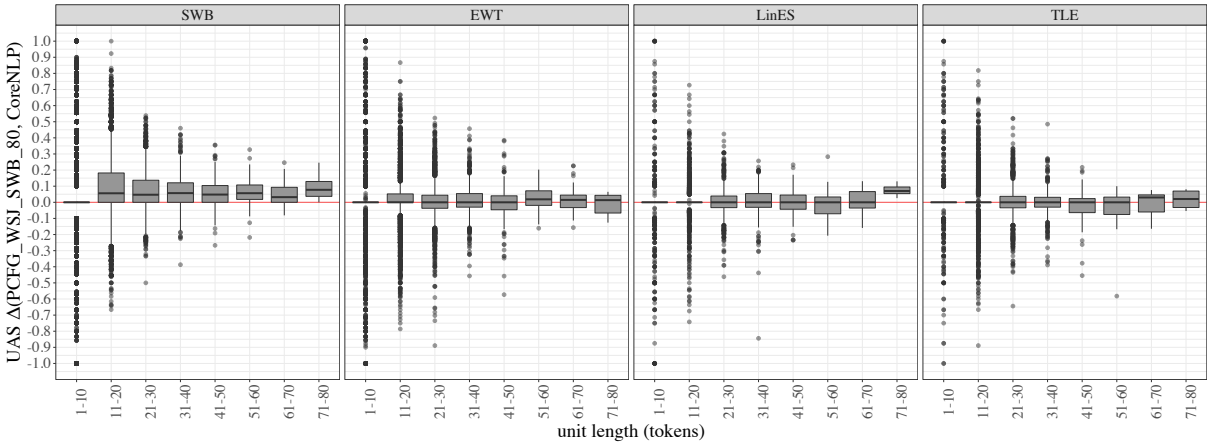

Figure 4: Unlabelled attachment scores by unit length in four English corpora: difference between model PCFG_WSJ_SWB_80 and CoreNLP.

maintained. We do not criticise the software in any way for deteriorating performance on long speech-units, as this is a hard problem, and we suspect that any other parser would suffer in similar ways. Indeed another option for future work is to use other publicly available parsers such as MSTParser (McDonald et al., 2006), TurboParser (Martins et al., 2013) and MaltParser (Nivre et al., 2007) to compare performance and potentially spot parsing errors through disagreement, per the method described by Smith & Dickinson (2014).

5 Conclusion and future work

In this paper we have shown that there are many differences between speech and writing at lexical and morphological levels. We also report how parser performance deteriorates as the input unit lengthens: an outcome which is perhaps unsurprising but which we showed to be especially acute for spoken language. Finally, we trained a new parsing model with added speech data and reported improvements for UAS scores across the board – more so for speech than writing. We make the models publicly available for other researchers[5] and welcome improved models or training data from others.

In future work we plan to analyse samples of speech-units with low UAS scores, to discover whether there are systematic parsing errors which could be solved through algorithmic changes to the parser, extra pre-processing steps, or otherwise. We also intend to continue comparing lexical and morpho-syntactic distributions in spoken and written corpora – dependency relations for example – to identify differences which may have implications for parsing. We suspect there may be lessons to be learned from parse tree analysis of learner text, such as the association between omission of the main verb and parse error (Ott and Ziai, 2010).

With more training data we can produce better parsing models, and potentially pursue self-learning algorithms in training. We might also introduce a heuristic to deal with long speech-units, which are particularly troublesome for existing parsers. One technique we can adopt is that of 'clause splitting', or 'chunking', which subdivides long strings for the purpose of higher quality analysis over small units (Tjong et al., 2001; Muszyńska, 2016). We hypothesise that such a step would play to the strength of existing parsers, namely their robustness over short inputs.

Acknowledgments

This paper reports on research supported by Cambridge English, University of Cambridge. We are grateful to our colleagues Calbert Graham and Russell Moore. We thank Sebastian Schuster and Chris Manning of Stanford University for their assistance with the Universal Dependencies corpora. We acknowledge the reviewers for their very helpful comments and have attempted to improve the paper in line with their suggestions.

[5] https://goo.gl/iQMu9w

References

Lars Ahrenberg. 2015. Converting an English-Swedish parallel treebank to Universal Dependencies. In *Proceedings of the Third International Conference on Dependency Linguistics (DepLing 2015)*.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for Learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English Web Treebank LDC2012T13.

David Brazil. 1995. *A grammar of speech*. Oxford: Oxford University Press.

Aoife Cahill, Binod Gyawali, and James Bruno. 2014. Self-training for parsing learner text. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*.

Andrew Caines and Paula Buttery. 2014. The effect of disfluencies and learner errors on the parsing of spoken learner language. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*.

Ronald Carter and Michael McCarthy. 2017. Spoken Grammar: where are we and where are we going? *Applied Linguistics* 38:1–20.

Eugene Charniak and Mark Johnson. 2001. Edit detection and parsing for transcribed speech. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Association for Computational Linguistics.

Joachim Daiber and Rob van der Goot. 2016. The Denoised Web Treebank: evaluating dependency parsing under noisy input conditions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of NAACL-HLT 2013*. Association for Computational Linguistics.

Jennifer Foster, Joachim Wagner, and Josef van Genabith. 2008. Adapting a WSJ-trained parser to grammatically noisy text. In *Proceedings of ACL-08: HLT, Short Papers*.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *Proceedings of Acoustics, Speech, and Signal Processing (ICASSP-92)*. IEEE.

Yoav Goldberg, Yuval Marton, Ines Rehbein, Yannick Versley, Özlem Çetinoğlu, and Joel Tetreault, editors. 2014. *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*. Association for Computational Linguistics.

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

Jeremy G. Kahn, Matthew Lease, Eugene Charniak, Mark Johnson, and Mari Ostendorf. 2005. Effective use of prosody in parsing conversational speech. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Jeremy G. Kahn and Mari Ostendorf. 2012. Joint reranking of parsing and word recognition with automatic segmentation. *Computer Speech and Language* 26(1):1–19.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency parsing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

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

Ann Lee and James Glass. 2012. Sentence detection using multiple annotations. In *Proceedings of INTERSPEECH 2012*. International Speech Communication Association.

Geoffrey Leech. 2000. Grammars of spoken English: new outcomes of corpus-oriented research. *Language Learning* 50:675–724.

William Levelt. 1989. *Speaking: from intention to articulation*. Cambridge, MA: MIT Press.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60.

Mitchell Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3 LDC99T42.

André Martins, Miguel Almeida, and Noah Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X)*.

Marie Meteer, Ann Taylor, Robert MacIntyre, and Rukmini Iyer. 1995. *Dysfluency annotation stylebook for the Switchboard corpus*. Philadelphia: Linguistic Data Consortium.

Seyed Mirroshandel and Alexis Nasr. 2011. Active learning for dependency parsing using partially annotated sentences. In *Proceedings of the 12th International Conference on Parsing Technologies*.

Russell Moore, Andrew Caines, Calbert Graham, and Paula Buttery. 2015. Incremental dependency parsing and disfluency detection in spoken learner English. In *Proceedings of the 18th International Conference on Text, Speech and Dialogue (TSD)*. Berlin: Springer-Verlag.

Russell Moore, Andrew Caines, Calbert Graham, and Paula Buttery. 2016. Automated speech-unit delimitation in spoken learner English. In *Proceedings of COLING*.

Ewa Muszyńska. 2016. Graph- and surface-level sentence chunking. In *Proceedings of the ACL 2016 Student Research Workshop*.

Alexis Nasr, Frederic Bechet, Benoit Favre, Thierry Bazillon, Jose Deulofeu, and Andre Valli. 2014. Automatically enriching spoken corpora with syntactic information for linguistic studies. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Marie Candito, Gülşen Cebirolu Eryiit, Giuseppe G. A. Celano, Fabricio Chalub, Jinho Choi, Çar Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökrmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Linh Hà M, Dag Haug, Barbora Hladká, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşkara, Hiroshi Kanayama, Jenna Kanerva, Natalia Kotsyba, Simon Krek, Veronika Laippala, Phng Lê Hng, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Lng Nguyn Th, Huyn Nguyn Th Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalnia, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2017. Universal Dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gulsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: a language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13:95–135.

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*.

Barbara Plank. 2016. What to do about non-standard (or *non-canonical*) language in NLP. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*.

Marwa Ragheb and Markus Dickinson. 2014. The effect of annotation scheme decisions on parsing learner data. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories*.

Mohammad S. Rasooli and Joel Tetreault. 2013. Joint parsing and disfluency detection in linear time.

In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Marek Rei and Ted Briscoe. 2013. Parser lexicalisation through self-learning. In *Proceedings of NAACL-HLT 2013*.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication* 32:127–154.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Amber Smith and Markus Dickinson. 2014. Evaluating parse error detection across varied conditions. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories*.

Stephanie Strassel. 2003. *Simple metadata annotation specification*. Version 5.0.

Erik Tjong, Kim Sang, and Herv'e Déjean. 2001. Introduction to the conll-2001 shared task: Clause identification. In *Proceedings of the 2001 Workshop on Computational Natural Language Learning (CoNLL)*.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*.

Masashi Yoshikawa, Hiroyuki Shindo, and Yuji Matsumoto. 2016. Joint transition-based dependency parsing and disfluency detection for automatic speech recognition texts. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.