The NAIST-NTT TED Talk Treebank

Graham Neubig[*], *Katsuhito Sudoh*[†], *Yusuke Oda*[*], *Kevin Duh*[*], *Hajime Tsukuda*[†], *Masaaki Nagata*[†]

[*] Nara Institute of Science and Technology, Nara, Japan
[†] NTT Communication Science Laboratories, Kyoto, Japan
neubig@is.naist.jp

Abstract

Syntactic parsing is a fundamental natural language processing technology that has proven useful in machine translation, language modeling, sentence segmentation, and a number of other applications related to speech translation. However, there is a paucity of manually annotated syntactic parsing resources for speech, and particularly for the lecture speech that is the current target of the IWSLT translation campaign. In this work, we present a new manually annotated treebank of TED talks that we hope will prove useful for investigation into the interaction between syntax and these speech-related applications. The first version of the corpus includes 1,217 sentences and 23,158 words manually annotated with parse trees, and aligned with translations in 26-43 different languages. In this paper we describe the collection of the corpus, and an analysis of its various characteristics.

1. Introduction

Syntactic parsing is widely considered as a useful component of natural language processing systems, not the least of which being machine translation [1, 2]. While a large part of the work on these applications has focused on the written word, we can assume that the fundamental principles behind syntax's success in these applications will also carry over to spoken language as well.

The great majority of recent work on syntactic parsing has been based on the statistical paradigm, in which the parameters of the parser are estimated from treebanks of manually annotated parse trees. In English, the standard data set for estimating these parsers is the Wall Street Journal section of the Penn Treebank [3], consisting of written language from newspapers. However, as there are large differences between written language and spoken language, there have also been some efforts to create resources for spoken language, including the Penn Treebank annotations of ATIS travel conversation and Switchboard telephone conversation data, as well as the OntoNotes [4] annotation of broadcast news and commentary. While these corpora mainly focus on informal speech or news, spoken monologue in the form of talks presented to an audience is also an attractive target for speech processing applications. In particular, the talk data from TED[1] has been used as a target for much research, most notably the IWSLT evaluation campaigns [5].

In this work, we present the *NAIST-NTT TED Talk Treebank*, a new manually annotated treebank of TED talks that we hope will prove useful for investigation into the interaction between syntax and speech-related applications such as speech translation. The first version of the corpus consists of a total of 10 talks, consisting of approximately 125 minutes of audio amounting to 1217 sentences. All sentences are manually annotated with parse trees following the standard Penn Treebank format. To allow for examination of the interaction between syntax and speech, all sentences are automatically time aligned with the corresponding speech file. In addition, to allow for multi-lingual research, we collected and sentence-aligned TED subtitles in anywhere from 26 to 43 languages per talk, with a total of 18 languages having translations for every talk.

In this paper, we present the details of how we constructed the corpus, including data collection, treebank annotation, speech time alignment, and multilingual sentence alignment. We also provide an analysis of the corpus, including its various characteristics and to what extent they differ from existing speech and text corpora, as well as the accuracy of an existing syntactic parser on the corpus. The corpus has been made publicly available for download under the Creative Commons License at
http://ahclab.naist.jp/resource/tedtreebank

2. Corpus Data

In this section, we describe the data used as material for the corpus.

2.1. English Data

Table 1: *Details of the annotated data.*

Set	Talk	Min.	Sent.	Word
All	10	125.07	1,217	23,158
Train	7	87.23	822	16,063
Test	3	37.84	395	7,095

The English text and speech data were gathered from TED Talks. Specifically, we gathered data starting with the beginning of

[1] http://www.ted.com

the May 2012 version of the WIT3 [6] training corpus for English-Japanese. From this data, for the first version of the treebank we chose 10 talks, the details of which are shown in Table 1.[2]

As the original TED data is subtitles, it is necessary to group these subtitles into sentences before performing annotation. In the creation of the corpus, we used the standard English sentence segmentation provided by the WIT3 data.[3]

In addition, when using a corpus for experiments, it is desirable to have a "standard" split between the training and testing data. As this standard, we designated a split of the first 7 talks as training data, and the other 3 talks as test data, resulting in an approximately 2/3 of the corpus for training, and 1/3 for testing when counting the number of sentences. This is also the split used in the analysis in Section 5.

With regards to the characteristics of the speeches and the speakers, the collected data is, like TED as a whole, quite diverse. Of the ten talks, 9 have a single speaker, and 1 has two speakers. Of these 11 speakers, 7 are men, and 4 are women.

2.2. Multilingual Data

In addition, because most of the talks in the collection have been translated into several other languages, we also downloaded the subtitles for all other languages in which they existed. As a result, for each talk we obtained subtitles in 26-43 different languages. For a total of 18 languages (shown in Table 2), this resulted in subtitles for all the parsed talks, and for 37 languages there were subtitles for some, but not all of the talks. We further combined these subtitles together into units that correspond to each English sentence, creating a sentence-aligned corpus between all of the languages.[4]

Table 2: *Languages for which subtitles existed for all 10 annotated talks.*

Arabic, Bulgarian, German, Greek, Spanish, French,
Hebrew, Italian, Japanese, Korean, Dutch, Polish,
Brazilian Portuguese, Romanian, Russian, Turkish,
Simplified Chinese, Traditional Chinese

While there exist other corpora of sentence-aligned TED talks [6], and other corpora of bilingually aligned syntax trees [7], to our knowledge this is the first corpus with manually annotated syntax trees in English and translations into a large number of languages, and also the first multilingually aligned treebank of the spoken word. We hope that this data will be of use for investigations into the effect of syntax on speech translation and other cross-lingual tasks.

3. Creation of Parse Trees

The first, and most labor-intensive annotation task was the creation of manual parse trees for the English sentences.

3.1. Annotation Standard

The most important part of creating a treebank is coming up with an appropriate annotation standard. Fortunately, the extensive 318-page annotation standard for the Penn Treebank exists,[5] and we choose to adopt this standard to maintain intercompatibility with the Penn Treebank. Specifically, we follow the actual documentation of the Treebank II annotation standard, but only annotate constituent labels (e.g. "NP"), omitting tagging of syntactic roles (e.g. the "-SUBJ" in "NP-SUBJ") or null elements (e.g. the omitted subject due to wh-movement in questions). We chose this annotation standard because most treebank parsers, such as the Berkeley parser, are trained on and generate annotation without constituent labels or null elements.

We also make one minor modification of the treebank standard tailored to the speech that appears in TED. Specifically, within TED talks, there are many cases in which the speaker quotes the words of another. The quote annotation in the Penn Treebank, in contrast to the annotation of other phenomena such as parenthesized expressions, simply treats each element of a quote as elements of its surrounding clause. In order to make the boundaries of quotes more explicit and easy to recognize, we add a single node with the symbol "QUOTE" showing the boundaries of a quote, as is done for parenthesized expressions. It should be noted that this change is automatically reversible, and the Penn Treebank annotation can be completely recovered by simply removing the QUOTE node and promoting its children.

An example of an annotated tree, including a QUOTE annotation is shown in Figure 1.

3.2. Annotation Process

Treebank annotation is an extremely time consuming process, particularly when the entirety of the tree has to be created from scratch. Fortunately, relatively accurate treebank parsers already exist, allowing us to create an initial parse first using an off-the-shelf parser, then have annotators spend their time fixing the errors of the existing parser. In this case, we use the Berkeley Parser[6] [8] to create an initial parse.

After this, we hired annotators to go through the trees and annotated them based on the standard described in the previous section. The annotators are well versed in annotation of linguistic data, and were given the standard and asked to follow it closely. After receiving this initial annotation result, the first author of the paper went through the entirety of the corpus, checking once more for any remaining errors. Finally, the trees were automatically checked for inconsistencies such as duplicated unary rules, or trees that were judged as a warning or error according to the phrase structure conversion tools of Johansson and Nugues [9].

4. Speech Time Alignment

Because the treebank described in this paper is of spoken language, the correspondence between syntactic trees and features of the speech is of particular interest. For example, it has been previously noted that prosody and syntax have a close relationship [10], and this corpus could be used to perform further investigations into these and other issues.

In order to create the time alignment of each word in the speech,

[2] We are currently in the process or annotating more data, which will be released as a second version of the corpus on completion.

[3] This segmentation standard groups multiple subtitles into single sentences, but never splits subtitles. Thus there are rare cases where a subtitle containing multiple sentences results in unsegmented sentences in the data.

[4] Of course, there are also a few cases where a single English sentence corresponds to multiple sentences, or less than one sentence in the foreign language.

[5] `http://www.cis.upenn.edu/~treebank/`

[6] `https://code.google.com/p/berkeleyparser/`

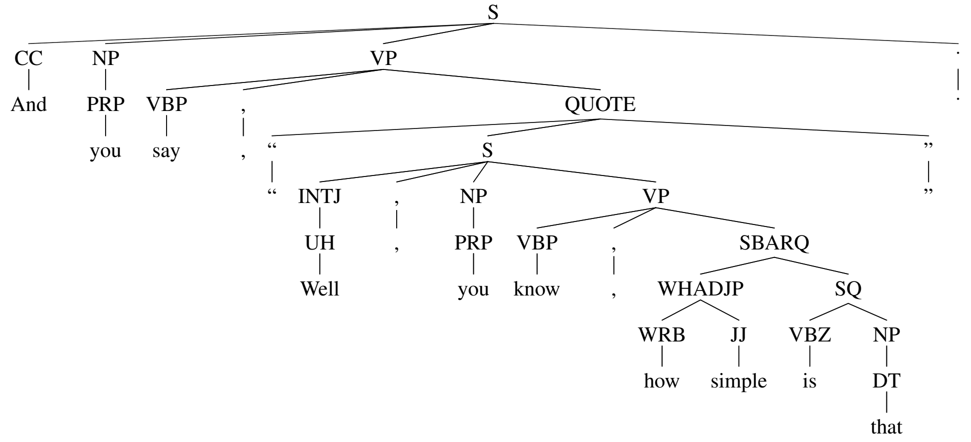

Figure 1: *An example tree from TED including QUOTE annotation.*

we prepared the data according to an automatic process. In the first step of the process, we performed forced decoding using the Kaldi decoder [11] with a model trained for the IWSLT speech recognition task [12]. In addition, as there are small differences between the transcripts used in forced decoding and the actual subtitles, due to factors such as punctuation deletion and normalization of numbers, we further aligned the times found in the forced alignment to the words in the subtitles, which were used in the annotation of the parse trees.

5. Analysis

In this section, we describe our analysis of the prepared corpus, first listing statistics of the trees in the corpus, measuring parsing accuracy and analyzing parsing errors.

5.1. Corpus Statistics

First, in this section we describe statistics of the collected parse trees for TED in comparison to the Wall Street Journal (WSJ) section of the Penn Treebank and the Broadcast News (BN) and Broadcast Commentary (BC) sections of OntoNotes. In particular, we focus on the differences in complexity of the sentences, as well as the different types of syntactic structures that appear in the sentences.

5.1.1. Syntactic Complexity

The first and most simple statistic that comes to mind regarding the complexity of the sentences is sentence length. In Figure 2 we show a histogram of the sentence lengths for the two corpora (after tokenization). From this figure we can see, perhaps as expected, that there is a larger number of long sentences in the newspaper text of WSJ. However, there are still a significant number of long sentences in TED with approximately 40% of sentences being 20 words or more. Compared with the two corpora of broadcast news and commentary, we can see that the length characteristics of the corpus are quite similar to those of broadcast news, and significantly longer than the more spontaneous broadcast commentary.

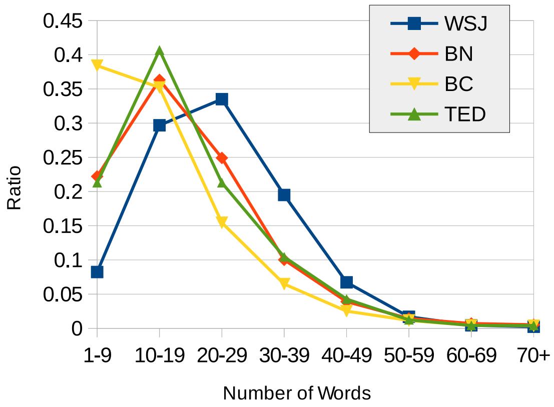

Figure 2: *A histogram of sentence lengths in Wall Street Journal (WSJ), Broadcast News (BN), Broadcast Commentary (BC), and TED.*

In addition to the length, it is also possible to examine the syntactic trees directly to understand the syntactic complexity of the sentences. There are a number of measures of syntactic complexity, and according to Roark et al. [13], who examine the correlation of several syntactic complexity measures with neuropsychological tests, two measures show a significant correlation with psychological factors such as the burden on memory. The first is simply the ratio of internal tree nodes to words in the sentence. The second is Frazier's measure of syntactic complexity [14], which is inspired by the number of syntactic elements that must be held in working memory. Specifically, it is defined as the average distance between a terminal node in the syntactic tree and its first ancestor that is not a leftmost sibling, with sentence nodes counting 1.5 times as much as other nodes (more details can be found in the referenced paper).

Table 3: *Syntactic complexity for sentences of length 10-29.*

Measure	WSJ	BN	BC	TED
Frazier	0.766	0.836	0.884	0.832
Nodes/Word	2.781	2.855	2.897	2.874

In Table 3 we show the values of these two complexity mea-

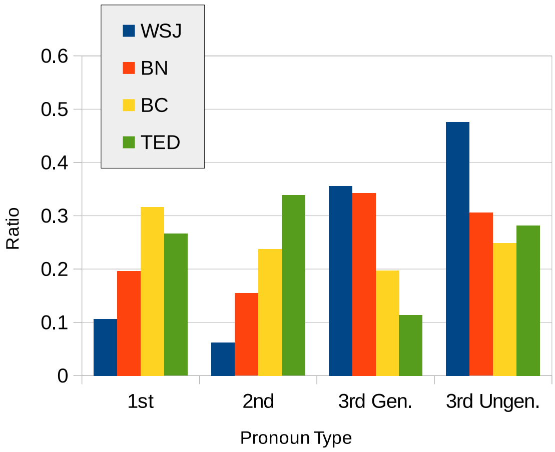

Figure 3: *The distribution of pronoun types for each corpus.*

sures for the 4 corpora under consideration, limiting our analysis to sentences of length 10-29 to reduce any artificial effects of analyzing different length sentences. From the results, we can see that WSJ has the lowest scores, BC has the highest scores, and TED and BN are relatively similar. While it seems somewhat counterintuitive that the more conversational corpora have more syntactic complexity, in fact news text is carefully planned and edited, often resulting in sentences that are easier to interpret than those in more informal speech.

5.1.2. *Stylistic Difference*

As the previous statistics show that the complexity of sentences in the TED corpus are similar to those of broadcast news, it is of interest whether there are stylistic differences that set it apart. It is somewhat difficult to pick apart stylistic differences quantitatively as simple statistics such as unigram distributions conflate stylistic and topical differences, so we calculated a variety of statistics and here focus on two simple statistics in which TED stood out.

First, in Figure 3, we show the difference in the distribution of singular pronouns, grouped into the first person (I/me), second person (you), third person gendered (he/she/him/her), and third person ungendered (it). From this figure, we can see that TED is unique in having more second person pronouns than any other category, demonstrating how TED speakers attempt to reference and engage their audience. In this way, the corpus is most similar to BC, which also contains a large number of 1st and 2nd person references, and in stark contrast to news, for which the large majority of pronouns are in the 3rd person.

Table 4: *Percentage of present, past, and progressive verbs.*

Tense	WSJ	BN	BC	TED
Present	42.8	50.2	56.1	64.0
Past	38.4	29.7	27.7	18.7
Prog.	18.7	20.1	16.1	17.3

Second, in Table 4, we show statistics about the tense of verbs, whether in the present (VBP/VBZ), past (VBD), or progressive (VBG) tense. From this table, we can see that as we move from news to conversation to TED, the number of past tense verbs decreases, and the number of present tense verbs increases. This marks a notable difference between news, which often looks backwards on the past, and the TED talks, which are often focused on what the speaker is doing now, or looking forward into the future.

In summary of the analysis, TED represents broadcast news in sentence complexity, but is also close to broadcast conversation in two stylistic characteristics. Thus, TED is somewhat different from these other genres, and thus manually annotated syntactic resources for TED are likely to give a benefit in the processing of TED talks and other similar monologues. In the following section, we examine this further in parsing experiments using the TED treebank.

5.2. Parsing Experiments

In order to test the accuracy of automatic parsing over the TED treebank, we performed parsing experiments, comparing with the WSJ section of the Penn Treebank.

5.2.1. *Experimental Setting and Accuracy*

We used two different sets of training data. The *wsj-train* data includes WSJ sections 2 to 21, which is the standard setting for training parsers on the Penn Treebank. The *wsj+ted-train* data also includes TED treebank training data (the first 7 talks, as specified in Section 2.1) in addition to *wsj-train*. We also prepared two data sets for testing each model. The *wsj-test* data includes WSJ section 23, the standard testing setting for evaluating syntactic parsers on WSJ. and *ted-test* data includes the TED treebank testing data (again specified in Section 2.1 as the last 3 talks). All “QUOTE” tags in the TED treebank are removed before training and testing, in order to ensure consistency with WSJ.

The Berkeley Parser [8] is used to train a latent annotated probabilistic context free grammar (PCFGLA) model from each of the training data sets and to generate a one-best parse of test data using trained model. We used EVALB[7] to evaluate parsing accuracy of each result in the form of bracketing F1 measure.

Table 5 shows the bracketing F1 measure for test sentences that have 40 words or less in each train/test data combination. Numbers in bold indicate the model with the better accuracy using the same test data. From these results, we can see that on the *ted-test* data, the model trained using the *wsj+ted-train* data achieves somewhat better performance than the model trained with only *wsj-train*. For *wsj-test*, the difference is slim, with both models achieving largely the same accuracy.

These results indicate that just by adding a small number of TED sentences to the WSJ data for training, we are able to achieve a small gain in parsing accuracy on the TED data. It should be noted that this is the simplest possible method for domain adaptation, and it is likely that there is still significant room for improvement by using more sophisticated techniques to account for the fact that the TED data is still significantly smaller than the WSJ data.

5.2.2. *Individual Examples*

Figures 4 and 5 show examples of parse trees of a sentence from the test set[8] trained with the *wsj-train* model, and the *wsj+ted* model respectively. The correct parse is the same as that generated by the *wsj+ted* model, with the exception that “(NN soap)” should be ”(NP

[7] `http://nlp.cs.nyu.edu/evalb/`

[8] The example was actually slightly shortened by removing two elements from the long coordinate phrase to ensure that it fit on one page.

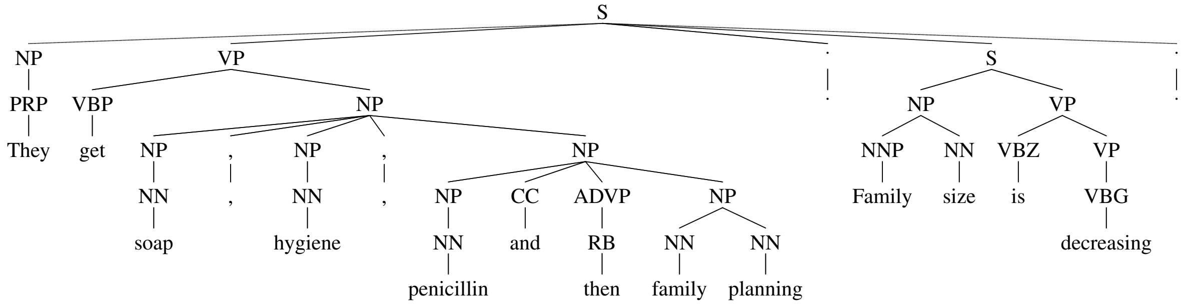
S
NP VP . S .
PRP VBP NP . NP VP .
They get NP , NP , NP NNP NN VBZ VP
NN , NN , NP CC ADVP NP Family size is VBG
soap hygiene NN and RB NN NN decreasing
penicillin then family planning

Figure 4: *Example of best parse using the* wsj-train *grammar.*

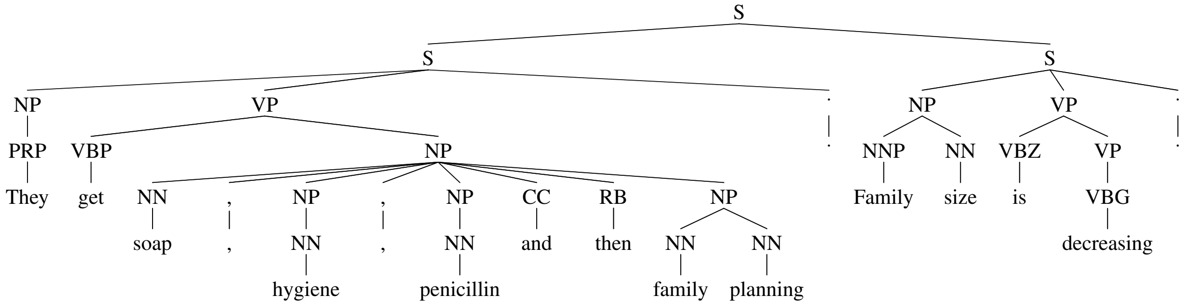
S
S S
NP VP . NP VP .
PRP VBP NP . NNP NN VBZ VP .
They get NN , NP , NP CC RB NP Family size is VBG
soap , NN , NN and then NN NN decreasing
hygiene penicillin family planning

Figure 5: *Example of the best parse using the* wsj+ted-train *grammar.*

Table 5: *Bracketing F1 measure of each parsing evaluation.*

		Train	
		wsj-train	*wsj+ted-train*
Test	*wsj-test*	**90.41**	90.38
	ted-test	88.65	**88.99**

(NN soap)),” and “(NNP Family)” should be “(NN Family).” This sentence has two notable characteristics.

First, there are multiple sentences in one tree, because TED data is based on subtitles of actual talks. These multiple sentence lines often occur when multiple sentences are included within a single subtitle. This is in contrast to the WSJ in which each line is split at sentence boundaries before annotation. As a result, the model trained using only the WSJ corpus tends to misparse lines including multiple sentences as single sentence.[9] On the other hand, model trained including the TED treebank expresses them using the (S → S S) rule and can parse sentences with this characteristic properly, although it does still make the mistake of determining that “Family” is a proper noun.

Second, the model trained by *wsj+ted-train* data makes a better parse of the long parallel noun phrase. In this example, the words “penicillin and then family planning” should be immediate children of the parent NP as in Figure 5, not an independent phrase as in Figure 4.

6. Conclusions

In this paper, we presented a treebank consisting of material from TED talks, an example of spoken language monologue sparsely covered by existing resources. The corpus consists of manually annotated syntactic trees, corresponding speech, time alignments, and multilingual translations. We hope that this corpus will be of use for examining the interaction between syntax and speech translation, or other applications of NLP to speech.

As future work, we are currently continuing annotation of the corpus, and plan to release an expanded second version of the corpus upon completion of this annotation. We also plan on performing more comprehensive parsing experiments using domain adaptation techniques, and examining the effect of parsing on the accuracy on machine translation.

7. References

[1] K. Yamada and K. Knight, “A syntax-based statistical translation model,” in *Proc. ACL*, 2001.

[2] G. Neubig and K. Duh, “On the elements of an accurate tree-to-string machine translation system,” in *Proc. ACL*, 2014.

[3] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, “Building a large annotated corpus of English: The Penn treebank,” *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[4] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel, “OntoNotes: the 90% solution,” in *Proc. HLT*, 2006, pp. 57–60.

[5] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, “Overview of the IWSLT 2012 evaluation campaign,” in *Proc. IWSLT*, 2012.

[9] A WSJ treebank parser with an extra sentence segmentation preprocessing step could also likely parse this example properly, but it this does add additional complexity that can be largely avoided by training a model that can handle these lines properly.

[6] M. Cettolo, C. Girardi, and M. Federico, “WIT3: web inventory of transcribed and translated talks,” 2012, pp. 261–268.

[7] N. Xue, F. Xia, F.-D. Chiou, and M. Palmer, “The Penn Chinese treebank: Phrase structure annotation of a large corpus,” *Natural language engineering*, vol. 11, no. 02, pp. 207–238, 2005.

[8] S. Petrov, L. Barrett, R. Thibaux, and D. Klein, “Learning accurate, compact, and interpretable tree annotation,” in *Proc. ACL*, 2006, pp. 433–440.

[9] R. Johansson and P. Nugues, “Extended constituent-to-dependency conversion for english,” in *16th Nordic Conference of Computational Linguistics*, 2007, pp. 105–112.

[10] S.-A. Jun, “Prosodic phrasing and attachment preferences,” *Journal of Psycholinguistic Research*, vol. 32, no. 2, pp. 219–249, 2003.

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, “The Kaldi speech recognition toolkit,” in *Proc. ASRU*, 2011, pp. 1–4.

[12] S. Sakti, K. Kubo, G. Neubig, T. Toda, and S. Nakamura, “The NAIST english speech recognition system for IWSLT 2013,” in *Proc. IWSLT*, 2013.

[13] B. Roark, M. Mitchell, and K. Hollingshead, “Syntactic complexity measures for detecting mild cognitive impairment,” in *Proc. BioNLP*, 2007, pp. 1–8.

[14] L. Frazier, “Syntactic complexity,” *Natural language parsing: Psychological, computational, and theoretical perspectives*, pp. 129–189, 1985.