Sentence segmentation of aphasic speech

Kathleen C. Fraser[1,3], **Naama Ben-David**[1], **Graeme Hirst**[1],
Naida L. Graham[2,3], **Elizabeth Rochon**[2,3]
[1]Department of Computer Science, University of Toronto, Toronto, Canada
[2]Department of Speech-Language Pathology, University of Toronto, Toronto, Canada
[3]Toronto Rehabilitation Institute, Toronto, Canada
{kfraser,naama,gh}@cs.toronto.edu, {naida.graham,elizabeth.rochon}@utoronto.ca

Abstract

Automatic analysis of impaired speech for screening or diagnosis is a growing research field; however there are still many barriers to a fully automated approach. When automatic speech recognition is used to obtain the speech transcripts, sentence boundaries must be inserted before most measures of syntactic complexity can be computed. In this paper, we consider how language impairments can affect segmentation methods, and compare the results of computing syntactic complexity metrics on automatically and manually segmented transcripts. We find that the important boundary indicators and the resulting segmentation accuracy can vary depending on the type of impairment observed, but that results on patient data are generally similar to control data. We also find that a number of syntactic complexity metrics are robust to the types of segmentation errors that are typically made.

1 Introduction

The automatic analysis of speech samples is a promising direction for the screening and diagnosis of cognitive impairments. For example, recent studies have shown that machine learning classifiers trained on speech and language features can detect, with reasonably high accuracy, whether a speaker has mild cognitive impairment (Roark et al., 2011), frontotemporal lobar degeneration (Pakhomov et al., 2010b), primary progressive aphasia (Fraser et al., 2014), or Alzheimer's disease (Orimaye et al., 2014; Thomas et al., 2005). These studies used manually transcribed samples of patient speech; however, it is clear that for such systems to be practical in the real world they must use automatic speech recognition (ASR). One issue that arises with ASR is the introduction of word recognition errors: insertions, deletions, and substitutions. This problem as it relates to impaired speech has been considered elsewhere (Jarrold et al., 2014; Fraser et al., 2013; Rudzicz et al., 2014), although more work is needed. Another issue, which we address here, is how ASR transcripts are divided into sentences.

turning to politics for al gore and george w bush another day of rehearsal in just over forty eight hours the two men will face off in their first of three debates for the first time voters will get a live unfiltered view of them together

Turning to politics, for Al Gore and George W Bush another day of rehearsal. In just over forty-eight hours the two men will face off in their first of three debates. For the first time, voters will get a live, unfiltered view of them together.

Figure 1: ASR text before and after processing.

The raw output from an ASR system is generally a stream of words, as shown in Figure 1. With some effort, it can be transformed into a format which is more readable by both humans and machines. Many algorithms exist for the segmentation of the raw text stream into sentences. However, there has been no previous work on how those algorithms might be applied to impaired speech.

This problem must be addressed for two reasons: first, sentence boundaries are important when analyzing the syntactic complexity of speech, which can be a strong indicator of potential impairment.

Many measures of syntactic complexity are based on properties of the syntactic parse tree (e.g. Yngve depth, tree height), which first require the demarcation of individual sentences. Even very basic measures of syntactic complexity, such as the mean length of sentence, require this information. Secondly, there are many reasons to believe that existing algorithms might not perform well on impaired speech, since assumptions about normal speech do not hold true in the impaired case. For example, in normal speech, pausing is often used to indicate a boundary between syntactic units, whereas in some types of dementia or aphasia a pause may indicate word-finding difficulty instead. Other indicators of sentence boundaries, such as prosody, filled pauses, and discourse markers, can also be affected by cognitive impairments (Emmorey, 1987; Bridges and Van Lancker Sidtis, 2013).

Here we explore whether we can apply standard approaches to sentence segmentation to impaired speech, and compare our results to the segmentation of broadcast news. We then extract syntactic complexity features from the automatically segmented text, and compare the feature values with measurements taken on manually segmented text. We assess which features are most robust to the noisy segmentation, and thus could be appropriate features for future work on automatic diagnostic interfaces.

2 Background

2.1 Automatic sentence segmentation

Many approaches to the problem of segmenting recognized speech have been proposed. One popular way of framing the problem is to treat it as a sequence tagging problem, where each interword boundary must be labelled as either a sentence boundary (B) or not (NB) (Liu and Shriberg, 2007).

Liu et al. (2005) showed that using a conditional random field (CRF) classifier for this problem resulted in a lower error rate than using a hidden Markov model or maximum entropy classifier. They stated that the CRF approach combined the benefits of these two other popular approaches, since it is discriminative, can handle correlated features, and uses a globally optimal sequence decoding.

The features used to train such classifiers fall broadly into two categories: word features and prosodic features. Word features can include word or part-of-speech *n*-grams, keyword identification, and filled pauses (Stevenson and Gaizauskas, 2000; Stolcke and Shriberg, 1996; Gavalda et al., 1997). Prosodic features include measures of pitch, energy, and duration of phonemes around the boundary, as well as the length of the silent pause between words (Shriberg et al., 2000; Wang et al., 2003).

The features which are most discriminative to the segmentation task can change depending on the nature of the speech. One important factor can be whether the speech is prepared or spontaneous. Cuendet et al. (2007) explored three different genres of speech: broadcast news, broadcast conversations, and meetings. They analyzed the effectiveness of different feature sets on each type of data. They found that pause features were the most discriminative across all groups, although the best results were achieved using a combination of lexical and prosodic features. Kolár et al. (2009) also looked at genre effects on segmentation, and found that prosodic features were more useful for segmenting broadcast news than broadcast conversations.

2.2 Primary progressive aphasia

There are many different forms of language impairment that could affect how sentence boundaries are placed in a transcript. Here, we focus on the syndrome of primary progressive aphasia (PPA). PPA is a form of frontotemporal dementia which is characterized by progressive language impairment without other notable cognitive impairment. In particular, we consider two subtypes of PPA: semantic dementia (SD) and progressive nonfluent aphasia (PNFA). SD is typically marked by fluent but empty speech, obvious word finding difficulties, and spared grammar (Gorno-Tempini et al., 2011). In contrast, PNFA is characterized by halting and sometimes agrammatic speech, reduced syntactic complexity, and relatively spared single-word comprehension (Gorno-Tempini et al., 2011). Because syntactic impairment, including reduced syntactic complexity, is a core feature of PNFA, we expect that measures of syntactic complexity would be important for a downstream screening application. Fraser et al. (2013) presented an automatic system for classifying PPA subtypes from ASR transcripts, but they were not able to include any syntactic complexity

metrics because their transcripts did not contain sentence boundaries.

3 Data

3.1 PPA data

Twenty-eight patients with PPA (11 with SD and 17 with PNFA) were recruited through three memory clinics, and 23 age- and education-matched healthy controls were recruited through a volunteer pool. All participants were native speakers of English, or had completed some of their education in English.

To elicit a sample of narrative speech, participants were asked to tell the well-known story of *Cinderella*. They were given a wordless picture book to remind them of the story; then the book was removed and they were asked to tell the story in their own words. This procedure, described in full by Saffran et al. (1989), is commonly used in studies of connected speech in aphasia.

The narrative samples were transcribed by trained research assistants. The transcriptions include filled pauses, repetitions, and false starts. Sentence boundaries were marked by a single annotator according to semantic, syntactic, and prosodic cues. We removed capitalization and punctuation, keeping track of original sentence boundaries for training and evaluation, to simulate a high-quality ASR transcript.

3.2 Broadcast news data

For the broadcast news data, we use a 804,064 word subset of the English section of the TDT4 Multilingual Broadcast News Speech Corpus[1]. Using the annotations in the transcripts, we extracted news stories only (ignoring teasers, miscellaneous text, and under-transcribed segments). The transcriptions were generated by closed captioning services and commercial transcription agencies (Strassel, 2005), and so they are of high but not perfect quality. Again, we remove capitalization and punctuation to simulate the output from an ASR system.

Since the TDT4 corpus is much larger than our PPA data set, we also construct a small news data set by randomly selecting 20 news stories from the TDT4 corpus. This allows us to determine which effects are due to differences in genre and which are due to having a smaller training set.

[1] `catalog.ldc.upenn.edu/LDC2005S11`

4 Methods

4.1 Lexical and POS features

The lexical features are simply the unlemmatized word tokens. We do not consider word *n*-grams due to the small size of our PPA data set. To extract our part-of-speech (POS) features, we first tag the transcripts using the NLTK POS tagger (Bird et al., 2009). We use the POS of the current word, the next word, and the previous word as features.

4.2 Prosodic features

To calculate the prosodic features, we first perform automatic alignment of the transcripts to the audio files. This provides us with a phone-level transcription, with the start and end of each phone linked to a time in the audio file. Using this information, we are able to calculate the length of the pauses between words, which we bin into three categories based on previous work by Pakhomov et al. (2010a). Each interword boundary either contains no pause, a short pause ($<$400 ms), or a long pause ($>$400 ms).

We calculate the pitch (Talkin, 1995; Brookes, 1997), energy, and duration of the last vowel before an interword boundary. For each measurement, we compare the value to the average value for that speaker, as well as to the values for the last vowel before the next and previous interword boundaries.

We perform the automatic alignment using the HTK toolkit (Young et al., 1997). Our pronunciation dictionary is based on the CMU dictionary[2], augmented with estimated pronunciations of out-of-vocabulary words using the "g2p" grapheme-to-phoneme toolkit (Bisani and Ney, 2008). We use a generic acoustic model that has been trained on Wall Street Journal text (Vertanen, 2006).

4.3 Classification

We use a conditional random field (CRF) to label each interword boundary as either a sentence boundary (B) or not (NB). We use a CRF implementation called CRFsuite (Okazaki, 2007) with the passive-aggressive learning algorithm. To avoid overfitting, we set the minimum feature frequency cut-off to 20.

To evaluate the performance of our system, we compare the hypothesized sentence boundaries with

[2] `www.speech.cs.cmu.edu/cgi-bin/cmudict`

the manually annotated sentence boundaries and report the F score, where F is the harmonic mean of recall and precision. For the PPA data and the small news data, we assess the system using a leave-one-out cross validation framework, in which each narrative is sequentially held out as test data while the system is trained on the remaining narratives. For the large TDT4 corpus, we randomly hold out 10% of the corpus as test data, and train on the remaining 90%.

4.4 Assessment of syntactic complexity

Once we have segmented the transcripts, we want to assess how the (presumably noisy) segmentation affects our measures of syntactic complexity. Here we consider a number of syntactic complexity metrics that have been previously used in the study of PPA speech (Fraser et al., 2014). The metrics are defined in the first column of Table 3. For the first four metrics, we generated the parse tree for each sentence using the Stanford parser (Klein and Manning, 2003). The Yngve depth is a well-known measure of how left-branching a parse tree is (Sampson, 1997; Yngve, 1960). The remaining metrics in Table 3 were calculated using Lu's Syntactic Complexity Analyzer (SCA) (Lu, 2010). We follow Lu's definitions for the various syntactic units: a *clause* is a structure consisting of at least a subject and a finite verb, a *dependent clause* is a clause which could not form a sentence on its own, a *verb phrase* is a phrase consisting of at least a verb and its dependents, a *complex nominal* is a noun phrase, clause, or gerund that stands in for a noun, a *coordinate phrase* is an adjective, adverb, noun, or verb phrase immediately dominated by a coordinating conjunction, a *T-unit* is a clause and all of its dependent clauses, and a *complex T-unit* is a T-unit which contains a dependent clause.

5 Segmentation results

5.1 Comparison between data sets

Table 1 shows the performance on the different data sets when trained using different combinations of feature types. We also report the chance baseline for comparison.

We first consider the differences in results observed between the two news data sets. The best results are similar in both groups, although, as would be expected, the larger training sample performs better. However, the difference is small, which suggests that the small size of the PPA data set should not greatly hurt the performance. When we compare the performance of these two groups with different sets of training features, we notice that the difference in performance is greatest when training on lexical features. In a small random sample from the TDT4 corpus, it is unlikely that two stories will cover the same topic, and so there will be little overlap in vocabulary. This is reflected in the results showing that lexical features hurt the performance in this small news sample.

Feature set	TDT4	TDT4 (small)	Controls	SD	PNFA
Chance baseline	0.07	0.07	0.05	0.07	0.06
All	**0.61**	0.57	**0.51**	**0.43**	**0.47**
Lexical+prosody	0.57	0.50	0.44	0.30	0.33
Lexical+POS	0.48	0.36	0.36	0.36	0.40
POS+prosody	**0.61**	**0.59**	0.45	0.39	0.45
POS	0.45	0.39	0.28	0.35	0.39
Prosody	0.50	0.48	0.24	0.23	0.25
Lexical	0.26	0.14	0.18	0.17	0.18

Table 1: F score for the automatic segmentation method on each data set. Boldface indicates best in column.

Performance on the news corpus is better than on the PPA data (including the control group). Comparing the small news sample to the PPA controls, we see that this is not simply due to the size of the training set, so we instead attribute the effect to the fact that speech in broadcast news is often prepared, while in the PPA data sets it is spontaneous.

A closer look at the effect of prosodic features in our training data further shows the difference we observe between prepared and spontaneous speech. When trained on the prosodic features alone, the news data set performs relatively well, while performance on the control data is much worse. These results are consistent with the findings of Kolár et al. (2009) regarding the effect of prosodic features in prepared and spontaneous speech.

When comparing the performance on the control group and on the PPA data, we see that generally, the results are better on the controls. This is to be expected, as the speech in the control group has more

complete sentences and fewer disfluencies. However, it is interesting to note that performance on the PNFA and SD groups is not much worse. All three data sets achieved the best results when trained with all feature types. This suggests that standard methods of sentence segmentation for spontaneous speech can be effective on PPA speech as well.

Looking at the PPA and control groups with other feature sets, we see that POS features are more important in the PNFA and SD groups than they are for the control data. A closer look at the transcripts shows us that the PPA participants tend to connect independent clauses with a conjunction more frequently than control participants, and independent clauses are often separated in the manual segmentation. This means that many sentence boundaries in the PPA data are marked by conjunctions. This is discussed further in the next section.

When considering the prosodic and lexical feature sets individually, we see that performance is similar in all three cases (control, SD, and PNFA). However, when we combine prosodic and lexical features together, the performance in the control case increases by a much larger margin than in the two aphasic cases. This suggests that control participants combine words and prosody in a manner that is more predictive of sentence boundaries than in the aphasic case.

TDT-4 (small)	**Control**	**SD**	**PNFA**
PRP_next	long pause	CC_next	long pause
DT_next	*go*	NNS	CC_next
RB	*her*	RB	NN
NNS	NNS	NN	RB_next
long pause	CC_next	RB_next	NNS
pitch<ave	RB	PRP_next	RB
NN	RB_next	energy<ave	short pause
CC_next	PRP_next	RB_prev	PRP_next
energy<ave	IN	VB	no pause
IN_prev	short pause	IN_prev	RB_prev

(a) Features associated with a boundary

TDT-4 (small)	**Control**	**SD**	**PNFA**
VBD_next	TO_next	*the*	TO_next
the	*so*	PRP$_next	*then*
IN	CC	*and*	*the*
MD_next	NNS_next	*then*	*she*
CC	*the*	VBD_next	VBP_next
VBG_next	*she*	VBZ_next	*and*
VBN_next	*and*	TO_next	*uh*
CD_prev	VBD_next	*'s*	VB_next
a	*of*	*I*	VBD_next
to	*uh*	*a*	*a*

(b) Features associated with a non-boundary

Table 2: The 10 features with the highest weights in each CRF model, indicating either that the following interword boundary is or is not a sentence boundary.

5.2 Important features

In Table 2, we report the 10 features in each data set which are most strongly associated with a boundary or a non-boundary. We consider only the small news corpus, for a fair comparison with the PPA data.

The POS tags shown are the output of the NLTK part of speech tagger, which uses the Penn Treebank Tag Set. We append '_next' and '_prev' to indicate that this is the POS tag of the next and previous word respectively. Italicized words represent lexical items.

We first consider the features that indicate a sentence boundary (see Table 2a). In general, we observe that our minimum frequency cut-off removes many of the lexical features from the top 10. (In the absence of such a cut-off, we observed that very low frequency words can be given deceptively high weights.) The exceptions to this are the words *go* and *her* in the control set. When we look at the data, there are indeed many occurrences of *go* and *her* at the end of sentences, for example, *she was not allowed to go* or *she couldn't go*, and *very mean to her* or *so in love with her*. While these lexical items are not specific to the *Cinderella* story, it seems unlikely that these features would generalize to other story-telling tasks (although we note that the Cinderella story is very widely used in the assessment of aphasia and some types of dementia).

The POS of the given word and its neighbours are generally important features. In all four cases, the next word being a coordinating conjunction or a pronoun is indicative of a boundary. In the three PPA cases, but not the news case, the next word being an adverb is also indicative. Looking at the data, we observe that this very often corresponds to the use of words like *so*, *then*, *well*, *anyway*, etc. This would seem to reflect a difference between the frequent use of discourse markers in spontaneous speech and their relative sparsity in prepared speech.

The POS of the current word is also important. In

all cases, a boundary is associated with the current word being an adverb or a noun. In the control data only, the tag IN, representing either a preposition or a subordinating clause, is also associated with a boundary. Although this seems counter-intuitive, an examination of the data reveals that in almost every case, this corresponds to the phrase *happily ever after*. The fact that this feature does not occur in the other PPA groups could indicate that the patients are less likely to use this phrase, but could also be due to our relatively high frequency cut-off.

Another anomalous result is that the tag VB (verb, base form) is associated with a sentence boundary in the SD case only. Again, examples from the data suggest a probable explanation. In many cases, sentences ending with VB are actually statements about the difficulty of the task, rather than narrative content; e.g., *that's all I can say*, *I can't recall*, or *I don't know*. These statements are consistent with the word-finding difficulties that are a hallmark of SD.

In the prosodic features, we see that long pauses and decreases in pitch and energy are associated with sentence boundaries in the news corpus. However, the results are mixed in the PPA data. This finding is consistent with our results in Section 5.1, and supports the conclusion of Cuendet et al. (2007) and Kolár et al. (2009) that prosodic features are more useful in prepared than spontaneous speech.

We now look briefly at the features which are associated with a non-boundary (Table 2b). Here we see more lexical features in the top 10, mostly function words and filled pauses. These features reflect the reasonable assumption that most sentences do not end with determiners, conjunctions, or subjective pronouns. One feature which occurs in the news data but not the PPA data is the next word being a modal verb (MD). This seems to be a result of the more frequent use of the future tense in the news stories (e.g. *the senator will serve another term*), in contrast to the Cinderella stories, which are generally told in the present or simple past tense.

6 Complexity results

We first compare calculating the syntactic complexity metrics on the manually segmented transcripts and the automatically segmented transcripts. The results are given in Table 3. Metrics for which there is no significant difference between the manual and automatic segmentation are marked with "NS". Of course, we do not claim that there is actually no difference between the values, as can be seen in the table, but we use this as a threshold to determine which features are less affected by the automatic segmentation.

All the features relating to Yngve depth and height of the parse trees are significantly different (in at least one of the three clinical groups). However, of the eight primary syntactic units calculated by Lu's SCA, six show no significant difference when measured on the automatically segmented transcripts. To examine this effect further, we will discuss how each of the eight is affected by the segmentation process.

Although the number of sentences (S) is different, the number of clauses (C) is not significantly affected by the automatic segmentation, which implies that the boundaries are rarely placed within clauses, but rather between clauses. An example of this phenomenon is given in Example 1:

Manual: And then they go off to the ball and then she comes I dunno how she meets up with this um fairy godmother whatever.

Auto: And then they go off to the ball. And then she comes I dunno how she meets up with this um fairy godmother whatever.

Our automatic method inserts a sentence boundary before the second *and*, breaking one sentence into two but not altering the number of clauses. In fact, the proposed boundary seems quite reasonable, although it does not agree with the human annotator. The correlation between the number of clauses counted in the manual and automatic transcripts is 0.99 in all three clinical groups. The counts for dependent clauses (DC) are also relatively unaffected by the automatic segmentation, for similar reasons.

The T-unit count (T) is also not significantly affected by the automatic segmentation. Since a T-unit only contains one independent clause as well as any attached dependent clauses, this suggests that the segmentation generally does not separate dependent clauses from their independent clauses. This also helps explain the lack of difference on complex T-units (CT).

Table 3 also indicates that the number of verb phrases (VP) and complex nominals (CN) is not significantly different in the automatically segmented

Metric	Diff?	Controls		SD		PNFA	
		Manual	Auto	Manual	Auto	Manual	Auto
Max YD maximum Yngve depth		5.10	4.53	4.45	3.87	4.66	3.83
Mean YD mean Yngve depth		2.97	2.72	2.68	2.44	2.77	2.41
Total YD total sum of the Yngve depths		66.92	53.41	42.48	32.95	49.95	32.57
Tree height average parse tree height		12.56	11.30	10.79	9.81	11.25	9.88
S number of sentences		24.35	31.22	27.73	37.36	18.82	25.47
T number of T-units	NS	31.43	35.13	32.55	39.27	23.29	27.41
C number of clauses	NS	61.48	64.48	57.73	62.45	42.94	46.65
DC number of dependent clauses	NS	24.70	27.30	26.27	26.09	16.59	18.88
CN number of complex nominals	NS	41.39	43.52	38.73	39.64	27.12	27.88
VP number of verb phrases	NS	77.00	79.65	72.09	77.00	51.76	55.24
CP number of coordinate phrases		12.39	10.30	11.55	6.91	7.82	4.18
CT number of complex T-units	NS	14.30	13.52	12.00	11.82	9.29	8.71
MLS mean length of sentence		19.79	16.22	14.04	11.25	15.86	11.60
MLT mean length of T-unit		14.92	13.72	12.19	10.46	12.78	10.66
MLC mean length of clause		7.55	7.21	7.13	6.58	6.89	6.39
T/S T-units per sentence		1.34	1.17	1.15	1.06	1.23	1.08
C/S clauses per sentence		2.64	2.25	1.96	1.70	2.28	1.82
DC/T dependent clauses per T-unit	NS	0.80	0.78	0.73	0.63	0.73	0.69
VP/T verb phrases per T-unit		2.47	2.34	2.11	1.92	2.23	1.98
CP/T coordinate phrases per T-unit		0.40	0.33	0.35	0.17	0.35	0.15
CN/T complex nominals per T-unit	NS	1.32	1.26	1.18	1.10	1.17	1.01
C/T clauses per T-unit		1.99	1.91	1.71	1.58	1.86	1.68
CT/T complex T-units per T-unit	NS	0.46	0.40	0.37	0.32	0.39	0.32
DC/C dependent clauses per clause	NS	0.39	0.41	0.42	0.40	0.38	0.41
CP/C coordinate phrases per clause		0.20	0.17	0.20	0.10	0.19	0.09
CN/C complex nominals per clause	NS	0.65	0.65	0.70	0.71	0.63	0.61

Table 3: Mean values of syntactic complexity metrics for the different patient groups. Features which show no significant difference between the manual and automatic segmentation on all three clinical groups are marked as "NS" (not significant).

transcripts. Since these syntactic units are typically sub-clausal, this is not unexpected given the arguments above.

The remaining primary syntactic unit, the coordinate phrase (CP), *is* different in the automatic transcripts. This represents a weakness of our method; namely, it has a tendency to insert a boundary before all coordinating conjunctions, as in Example 2:

Manual: So she is very upset and she's crying and with her fairy godmother who then uh creates a carriage and horses and horsemen and and driver and beautiful dress and magical shoes.

Auto: So she is very upset. And she's crying and with her. Fairy godmother who then uh creates a carriage. And horses and horsemen and and driver. And beautiful dress. And magical shoes.

In this case, the manual transcript has five coordinate phrases, while the automatic transcript has only two.

The mean lengths of sentence (MLS), clause (MLC), and T-unit (MLT) are all significantly different in the automatically segmented transcripts. We ascribe this to the fact that a small change in C or T can lead to a large change in MLC or MLT. The remaining metrics in Table 3 are simply combinations of the primary units discussed above.

Our analysis so far suggests that some syntactic units are relatively impervious to the automatic sentence segmentation, while others are more susceptible to error. However, when we examine the mean values given in Table 3, we observe that even in cases when the complexity metrics are significantly different in the automatic transcripts, the differences appear to be systematic. For example, we know that our segmentation method tends to produce more sentences than appear in the manual transcripts (i.e., S is always greater in the automatic transcripts). If we look at the differences across clinical groups, the same pattern emerges in both the man-

Metric	SD vs controls		PNFA vs controls		SD vs PNFA	
	manual	auto	manual	auto	manual	auto
Max YD	*	*	*	*		
Mean YD	*		*	*		
Total YD	*	*	*	*		
Tree height	*	*	*	*		
S						
T			*	*		
C			*	*		
DC			*	*		
CN			*	*		
VP			*	*		
CP			*	*		
CT			*	*		
MLS	*	*	*	*		
MLT	*	*	*	*		
MLC		*	*	*		
T/S	*					
C/S	*	*	*		*	
DC/T						
VP/T	*			*		
CP/T		*		*		
CN/T				*		
C/T	*					
CT/T	*	*				
DC/C						
CP/C		*		*		
CN/C						

Table 4: Difference between syntactic complexity metrics for each pair of patient groups. A significant difference ($p < 0.05$) is marked with an asterisk.

ual and automatic transcripts: participants with SD produce the most sentences, followed by controls, followed by participants with PNFA. In most applications of these syntactic complexity metrics, what matters most is not the absolute value of the metric, but the relative differences between groups. So, we now ask: which features distinguish between clinical groups in the manually segmented transcripts, and do they still distinguish between the groups in the automatically segmented transcripts?

Our results for this experiment are reported in Table 4. In the case of SD vs. controls, there are 11 features which are significantly different between the two groups in the manual transcripts. Seven of these features are also significantly different between groups in the automatic transcripts, while an additional three features are significant only in the automatic transcripts. In the PNFA vs. controls case, there are 15 distinguishing features in the manual transcripts, and 14 of those are also significantly different in the automatic transcripts. There are four features which are significant only in the automatic case. Finally, in the case of SD vs. PNFA, there is only one distinguishing feature in the manual transcripts, and none in the automatic transcript.

These findings suggest that automatically segmented transcripts can still be useful, even if the complexity metrics have different values from the manual transcripts. Importantly, a comparison of the mean feature values in Table 3 reveals that in 92% of cases, and in every case marked as significant in Table 4 4, the direction of the trend is the same in the manual and automatic transcripts.

For example, in the first column of Table 4, maximum Yngve depth is significantly different between SD participants and controls. In both the manual and automatic transcripts, the controls have a greater maximum depth than SD participants. This is true for every metric that is significant in both the manual and automatic transcripts. This indicates that the metrics are not only significantly different, and therefore useful for machine learning classification or some other downstream application, but that they are interpretable in relation to the specific language impairments that we expect to observe in the patient groups.

7 Discussion

We have introduced the issue of sentence segmentation of impaired speech, and tested the effectiveness of standard segmentation methods on PPA speech samples. We found that, as expected, performance was best on prepared speech from broadcast news, then on healthy controls, and worst on speech samples from PPA patients. However, the results on the PPA data are promising, and suggest that similar methods could be effective for impaired speech. Future work will look at adapting the standard algorithms to improve performance in the impaired case. This would include an evaluation of the forced alignment on impaired speech data, as well as the exploration of new features for the boundary classification.

One limitation of this study is the use of manually transcribed data with capitalization and punctuation removed to simulate perfect ASR data. We expect that real ASR data will contain recognition errors, and it is not clear how these errors will affect the segmentation process. As well, our PPA data set is relatively small from a machine learning perspec-

tive, due to the inherent difficulties associated with collecting clinical data. Furthermore, we assumed that the diagnostic group is known *a priori*, allowing us to train and test on each group separately.

We analyzed our results to see how the noise introduced by our segmentation affects various syntactic complexity measures. Some measures (e.g. T-units) were robust to the noise, while others (e.g. Yngve depth) were not. When using such automatic methods for the analysis of speech data, researchers should be aware of the unequal effects on different complexity metrics.

For the more practical goal of distinguishing between different patient groups, we found that most measures that were significant for this task using the manual transcripts remained so when using the automatically segmented ones, and the direction of the difference was the same in the manual and automatic transcripts. In all cases where a significant difference between the groups was detected, the direction of the difference was the same in the manual and automatic transcripts. These results indicate that imperfect segmentation methods might still be useful for some applications, since they affect the data in a systematic way.

Although we evaluated our methods against human-annotated data, there is some uncertainty about whether a single gold standard for the sentence segmentation of speech truly exists. Miller and Weinert (1998), among others, argue that the concept of a sentence as defined in written language does not necessarily exist in spoken language. In future work, it would useful to compare the inter-annotator agreement between trained human annotators to determine an upper bound for the accuracy.

Acknowledgments

Many thanks to Bruna Seixas Lima for providing the manual annotations. This work was supported by the Canadian Institutes of Health Research (CIHR), Grant #MOP-82744, and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. " O'Reilly Media, Inc.".

Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451.

Kelly Ann Bridges and Diana Van Lancker Sidtis. 2013. Formulaic language in Alzheimer's disease. *Aphasiology*, pages 1–12.

Michael Brookes. 1997. Voicebox: Speech processing toolbox for Matlab. `www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`.

Sebastien Cuendet, Elizabeth Shriberg, Benoit Favre, James Fung, and Dilek Hakkani-Tür. 2007. An analysis of sentence segmentation features for broadcast news, broadcast conversations, and meetings. *Searching Spontaneous Conversational Speech*, pages 43–49.

Karen D. Emmorey. 1987. The neurological substrates for prosodic aspects of speech. *Brain and Language*, 30(2):305–320.

Kathleen Fraser, Frank Rudzicz, Naida Graham, and Elizabeth Rochon. 2013. Automatic speech recognition in the diagnosis of primary progressive aphasia. In *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*, pages 47–54.

Kathleen C. Fraser, Jed A. Meltzer, Naida L. Graham, Carol Leonard, Graeme Hirst, Sandra E. Black, and Elizabeth Rochon. 2014. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, 55:43–60.

Marsal Gavalda, Klaus Zechner, et al. 1997. High performance segmentation of spontaneous speech using part of speech and trigger word information. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 12–15. Association for Computational Linguistics.

M. L. Gorno-Tempini, A. E. Hillis, S. Weintraub, A. Kertesz, M. Mendez, S. F. Cappa, J. M. Ogar, J. D. Rohrer, S. Black, B. F. Boeve, F. Manes, N. F. Dronkers, R. Vandenberghe, K. Rascovsky, K. Patterson, B. L. Miller, D. S. Knopman, J. R. Hodges, M. M. Mesulam, and M. Grossman. 2011. Classification of primary progressive aphasia and its variants. *Neurology*, 76:1006–1014.

William Jarrold, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria Luisa Gorno-Tempini, and Jennifer Ogar. 2014. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*, pages 27–36.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430.

Jáchym Kolár, Yang Liu, and Elizabeth Shriberg. 2009. Genre effects on automatic sentence segmentation of speech: A comparison of broadcast news and broadcast conversations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4701–4704. IEEE.

Yang Liu and Elizabeth Shriberg. 2007. Comparing evaluation metrics for sentence boundary detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 182–185. IEEE.

Yang Liu, Andreas Stolcke, Elizabeth Shriberg, and Mary Harper. 2005. Using conditional random fields for sentence boundary detection in speech. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 451–458, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Jim Miller and Regina Weinert. 1998. *Spontaneous spoken language*. Clarendon Press.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of conditional random fields (CRFs).

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen Jennifer Golden. 2014. Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. In *Proceedings of the 1st Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pages 78–87, Baltimore, Maryland. Association for Computational Linguistics.

S. V. Pakhomov, G. E. Smith, D. Chacon, Y. Feliciano, N. Graff-Radford, R. Caselli, and D. S. Knopman. 2010a. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and Behavioral Neurology*, 23:165–177.

Serguei V.S. Pakhomov, Glen E. Smith, Susan Marino, Angela Birnbaum, Neill Graff-Radford, Richard Caselli, Bradley Boeve, and David D. Knopman. 2010b. A computerized technique to assess language use patterns in patients with frontotemporal dementia. *Journal of Neurolinguistics*, 23:127–144.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffery Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2081–2090.

Frank Rudzicz, Rosalie Wang, Momotaz Begum, and Alex Mihailidis. 2014. Speech recognition in Alzheimer's disease with personal assistive robots. In *Proceedings of the 5th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 20–28. Association for Computational Linguistics.

Eleanor M. Saffran, Rita Sloan Berndt, and Myrna F. Schwartz. 1989. The quantitative analysis of agrammatic production: Procedure and data. *Brain and Language*, 37(3):440–479.

Geoffrey Sampson. 1997. Depth in English grammar. *Journal of Linguistics*, 33:131–51.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1):127–154.

Mark Stevenson and Robert Gaizauskas. 2000. Experiments on sentence boundary detection. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 84–89. Association for Computational Linguistics.

Andreas Stolcke and Elizabeth Shriberg. 1996. Automatic linguistic segmentation of conversational speech. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, volume 2, pages 1005–1008. IEEE.

Stephanie Strassel, 2005. *Topic Detection and Tracking Annotation Guidelines: Task Definition to Support the TDT2002 and TDT2003 Evaluations in English, Chinese and Arabic*. Linguistic Data Consortium, 1.5 edition.

David Talkin. 1995. A robust algorithm for pitch tracking (RAPT). *Speech Coding and synthesis*, 495:495–518.

Calvin Thomas, Vlado Keselj, Nick Cercone, Kenneth Rockwood, and Elissa Asp. 2005. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In *Proceedings of the IEEE International Conference on Mechatronics and Automation*, pages 1569–1574.

Keith Vertanen. 2006. Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments. Technical report, Cavendish Laboratory, University of Cambridge.

Dong Wang, Lie Lu, and Hong-Jiang Zhang. 2003. Speech segmentation without speech recognition. In *Proceedings of the IEEE International Conference Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 468–471. IEEE.

Victor Yngve. 1960. A model and hypothesis for language structure. *Proceedings of the American Physical Society*, 104:444–466.

Steve Young, Gunnar Evermann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland. 1997. *The HTK book*, volume 2. Entropic Cambridge Research Laboratory Cambridge.