Using statistical parsing to detect agrammatic aphasia

Kathleen C. Fraser[1], **Graeme Hirst**[1], **Jed A. Meltzer**[2],
Jennifer E. Mack[3], **and Cynthia K. Thompson**[3,4,5]
[1]Dept. of Computer Science, University of Toronto
[2]Rotman Research Institute, Baycrest Centre, Toronto
[3]Dept. of Communication Sciences and Disorders, Northwestern University
[4]Dept. of Neurology, Northwestern University
[4]Cognitive Neurology and Alzheimer's Disease Center, Northwestern University
{kfraser,gh}@cs.toronto.edu, jmeltzer@research.baycrest.org
{jennifer-mack-0,ckthom}@northwestern.edu

Abstract

Agrammatic aphasia is a serious language impairment which can occur after a stroke or traumatic brain injury. We present an automatic method for analyzing aphasic speech using surface level parse features and context-free grammar production rules. Examining these features individually, we show that we can uncover many of the same characteristics of agrammatic language that have been reported in studies using manual analysis. When taken together, these parse features can be used to train a classifier to accurately predict whether or not an individual has aphasia. Furthermore, we find that the parse features can lead to higher classification accuracies than traditional measures of syntactic complexity. Finally, we find that a minimal amount of pre-processing can lead to better results than using either the raw data or highly processed data.

1 Introduction

After a stroke or head injury, individuals may experience aphasia, an impairment in the ability to comprehend or produce language. The type of aphasia depends on the location of the lesion. However, even two patients with the same type of aphasia may experience different symptoms. A careful analysis of narrative speech can reveal specific patterns of impairment, and help a clinician determine whether an individual has aphasia, what type of aphasia it is, and how the symptoms are changing over time.

In this paper, we present an automatic method for the analysis of one type of aphasia, *agrammatic aphasia*. characterized by the omission of function words, the omission or substitution of morphological markers for person and number, the absence of verb inflection, and a relative increase in the number of nouns and decrease in the number of verbs (Bastiaanse and Thompson, 2012). There is often a reduction in the variety of different syntactic structures used, as well as a reduction in the complexity of those structures (Progovac, 2006). There may also be a strong tendency to use the canonical word order of a language, for example subject-verb-object in English (Progovac, 2006).

Most studies of narrative speech in agrammatic aphasia are based on manually annotated speech transcripts. This type of analysis can provide detailed and accurate information about the speech patterns that are observed. However, it is also very time consuming and requires trained transcribers and annotators. Studies are necessarily limited to a manageable size, and the level of agreement between annotators can vary.

We propose an automatic approach that uses information from statistical parsers to examine properties of narrative speech. We extract context-free grammar (CFG) production rules as well as phrase-level features from syntactic parses of the speech transcripts. We show that this approach can detect many features which have been previously reported in the aphasia literature, and that classification of agrammatic patients and controls can be achieved with high accuracy.

We also examine the effects of including speech dysfluencies in the transcripts. Dysfluencies and non-narrative words are usually removed from the transcripts as a pre-processing step, but we show that by retaining some of these items, we can actually achieve a higher classification accuracy than by using the completely clean transcripts.

Finally, we investigate whether there is any benefit to using the parse features instead of more traditional measures of syntactic complexity, such as Yngve depth or mean sentence length. We find that the parse features convey more information

Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014), pages 134–142,
Baltimore, Maryland USA, June 26-27 2014.

about the specific syntactic structures being produced (or avoided) by the agrammatic speakers, and lead to better classification accuracies.

2 Related Work

2.1 Syntactic analysis of agrammatic narrative speech

Much of the previous work analyzing narrative speech in agrammatic aphasia has been performed manually. One widely used protocol is called Quantitative Production Analysis (QPA), developed by Saffran et al. (1989). QPA can be used to measure morphological content, such as whether determiners and verb inflections are produced in obligatory contexts, as well as structural complexity, such as the number of embedded clauses per sentence. Subsequent studies have found a number of differences between normal and agrammatic speech using QPA (Rochon et al., 2000). Another popular protocol called the Northwestern Narrative Language Analysis (NNLA) was introduced by Thompson et al. (1995). This protocol analyzes each utterance at five different levels, and focuses in particular on the production of verbs and verb argument structure.

Perhaps more analogous to our work here, Goodglass et al. (1994) conducted a detailed examination of the syntactic constituents used by aphasic patients and controls. In that study, utterances were grouped according to how many syntactic constituents they contained. They found that agrammatic participants were more likely to produce single-constituent utterances, especially noun phrases, and less likely to produce subordinate clauses. They also found that agrammatic speakers sometimes produced two-constituent utterances consisting of only a subject and object, with no verb. This pattern was never observed in control speech.

A much smaller body of work explores the use of computational techniques to analyze agrammatism. Holmes and Singh (1996) analyzed conversational speech from aphasic speakers and controls. Their features mostly included measures of vocabulary richness and frequency counts of various parts-of-speech (e.g. nouns, verbs); however they also measured "clause-like semantic unit rate". This feature was intended to measure the speaker's ability to cluster words together, although it is not clear what the criteria for segmenting clause-like units were or whether it was done manually or automatically. Nonetheless, it was found to be one of the most important variables for distinguishing between patients and controls.

MacWhinney et al. (2011) presented several examples of how researchers can use the AphasiaBank[1] database and associated software tools to conduct automatic analyses (although the transcripts are first hand-coded for errors by experienced speech-language pathologists). Specifically with regards to syntax, they calculated several frequency counts and ratios for different parts-of-speech and bound morphemes. There was one extension beyond treating each word individually: this involved searching for pre-defined collocations such as *once upon a time* or *happily ever after*, which were found to occur more rarely in the patient transcripts than in the control transcripts.

We present an alternative, automated method of analysis. We do not attempt to fully replicate the results of the manual studies, but rather provide a complementary set of features which can indicate grammatic abnormalities. Unlike previous computational studies, we attempt to move beyond single-word analysis and examine which patterns of syntax might indicate agrammatism.

2.2 Using parse features to assess grammaticality

Syntactic complexity metrics derived from parse trees have been used by various researchers in studies of mild cognitive impairment (Roark et al., 2011), autism (Prud'hommeaux et al., 2011), and child language development (Sagae et al., 2005; Hassanali et al., 2013). Here we focus specifically on the use of CFG production rules as features.

Using the CFG production rules from statistical parsers as features was first proposed by Baayen et al. (1996), who applied the features to an authorship attribution task. More recently, similar features have been widely used in native language identification (Wong and Dras, 2011; Brooke and Hirst, 2012; Swanson and Charniak, 2012). Perhaps most relevant to the task at hand, CFG productions as well as other parse outputs have proved useful for judging the grammaticality and fluency of sentences. For example, Wong and Dras (2010) used CFG productions to classify sentences from an artificial error corpus as being either grammatical or ungrammatical.

Taking a different approach, Chae and Nenkova

[1] http://talkbank.org/AphasiaBank/

	Agrammatic ($N = 24$)	Control ($N = 15$)
Male/Female	15/9	8/7
Age (years)	58.1 (10.6)	63.3 (6.4)
Education (years)	16.3 (2.5)	16.4 (2.4)

Table 1: Demographic information. Numbers are given in the form: mean (standard deviation).

(2009) calculated several surface features based on the output of a parser, such as the length and relative proportion of different phrase types. They used these features to distinguish between human and machine translations, and to determine which of a pair of translations was the more fluent. However, to our knowledge there has been no work using parser outputs to assess the grammaticality of speech from individuals with post-stroke aphasia.

3 Data

3.1 Participants

This was a retrospective analysis of data collected by the the Aphasia and Neurolinguistics Research Laboratory at Northwestern University. All agrammatic participants had experienced a stroke at least 1 year prior to the narrative sample collection. Demographic information for the participants is given in Table 1. There is no significant ($p < 0.05$) difference between the patient and control groups on age or level of education.

3.2 Narrative task

To obtain a narrative sample, the participants were asked to relate the well-known fairy tale *Cinderella*. Each participant was first given a wordless picture book of the story to look through. The book was then removed, and the participant was asked to tell the story in his or her own words. The examiner did not interrupt or ask questions.

The narratives were recorded and later transcribed following the NNLA protocol. The data was segmented into utterances based on syntactic and prosodic cues. Filled pauses, repetitions, false starts, and revisional phrases (e.g. *I mean*) were all placed inside parentheses. The average length of the raw transcripts was 332 words for agrammatic participants and 387 words for controls; when the non-narrative words were excluded the average length was 194 words for the agrammatic group and 330 for controls.

4 Methods

4.1 Parser Features

We consider two types of features: CFG production rules and phrase-level statistics. For the CFG production rules, we use the Charniak parser (Charniak, 2000) trained on Wall Street Journal data to parse each utterance in the transcript and then extract the set of non-lexical productions. The total number of types of productions is large, many of them occurring very infrequently, so we compile a list of the 50 most frequently occurring productions in each of the two groups (agrammatic and controls) and use the combined set as the set of features. The feature values can be binary (does a particular production rule appear in the narrative or not?) or integer (how many times does a rule occur?). The CFG non-terminal symbols follow the Penn Treebank naming conventions.

For our phrase-level statistics, we use a subset of the features described by Chae and Nenkova (2009), which are related to the incidence of different phrase types. We consider three different phrase types: noun phrases, verb phrases, and prepositional phrases. These features are defined as follows:

- *Phrase type proportion*: Length of each phrase type (including embedded phrases), divided by total narrative length.
- *Average phrase length*: Total number of words in a phrase type, divided by number of phrases of that type.
- *Phrase type rate*: Number of phrases of a given type, divided by total narrative length.

Because we are judging the grammaticality of the entire narrative, we normalize by narrative length (rather than sentence length, as in Chae and Nenkova's study). These features are real-valued.

We first perform the analysis on the transcribed data with the dysfluencies removed, labeled the "clean" dataset. This is the version of the transcript that would be used in the manual NNLA analysis. However, it is the result of human effort and expertise. To test the robustness of the system on data that has not been annotated in this way, we also use the "raw" dataset, with no dysfluencies removed (i.e. including everything inside the parentheses), and an "auto-cleaned" dataset, in which filled pauses are automatically removed from the raw transcripts. We also use a simple algorithm to remove "stutters" and false starts, by

removing non-word tokens of length one or two (e.g. *C- C- Cinderella* would become simply *Cinderella*). This provides a more realistic view of the performance of our system on real data. We also hypothesize that there may be important information to be found in the dysfluent speech segments.

4.2 Feature weighting and selection

We assume that some production rules will be more relevant to the classification than others, and so we want to weight the features accordingly. Using *term frequency–inverse document frequency* (*tf-idf*) would be one possibility; however, the *tf-idf* weights do not take into account any class information. *Supervised term weighting* (STW), has been proposed by Debole and Sebastiani (2004) as an alternative to *tf-idf* for text classification tasks. In this weighting scheme, feature weights are assigned using the same algorithm that is used for feature selection. For example, one way to select features is to rank them by their information gain (InfoGain). In STW, the InfoGain value for each feature is also used to replace the *idf* term. This can be expressed as $W(i,d) = \mathrm{df}(i,d) \times \mathrm{InfoGain}(i)$, where $W(i,d)$ is the weight assigned to feature i in document d, $\mathrm{df}(i,d)$ is the frequency of occurrence of feature i in document d, and $\mathrm{InfoGain}(i)$ is the information gain of feature i across all the training documents.

We considered two different methods of STW: weighting by InfoGain and weighting by gain ratio (GainRatio). The methods were also used as feature selection, since any feature that was assigned a weight of zero was removed from the classification. We also consider *tf-idf* weights and unweighted features for comparison.

4.3 Syntactic complexity metrics

To compare the performance of the parse features with more-traditional syntactic complexity metrics (SC metrics), we calculate the mean length of utterance (MLU), mean length of T-unit[2] (MLT), mean length of clause (MLC), and parse tree height. We also calculate the mean, maximum, and total Yngve depth, which measures the proportion of left-branching to right-branching in each parse tree (Yngve, 1960). These measures are commonly used in studies of impaired language (e.g. Roark et al. (2011), Prud'hommeaux et al. (2011), Fraser et al. (2013b)). We hypothesize that the parse features will capture more information about the specific impairments seen in agrammatic aphasia; however, using the general measures of syntactic complexity may be sufficient for the classifiers to distinguish between the groups.

[2] A T-unit consists of a main clause and its attached dependent clauses.

4.4 Classification

To test whether the features can effectively distinguish between the agrammatic group and controls, we use them to train and test a machine learning classifier. We test three different classification algorithms: naive Bayes (NB), support vector machine (SVM), and random forests (RF). We use a leave-one-out cross-validation framework, in which one transcript is held out as a test set, and the other transcripts form the training data. The feature weights are calculated on the training set and then applied to the test set (as a result, each fold of training/testing may use different features and feature weights). The SVM and RF algorithms are tuned in a nested cross-validation loop. The classifier is then tested on the held-out point. This procedure is repeated across all data points, and the average accuracy is reported.

A baseline classifier which assigns all data to the largest class would achieve an accuracy of .62 on this classification task. For a more realistic measure of performance, we also compare our results to the baseline accuracy that can be achieved using only the length of the narrative as input.

5 Results

5.1 Features using clean transcripts

We first present the results for the clean transcripts. Although different features may be selected in each fold of the cross-validation, for simplicity we show only the feature rankings on the whole data set. Table 2 shows the top features as ranked by GainRatio. The frequencies are given to indicate the direction of the trend; they represent the average frequency per narrative for each class (agrammatic = AG and control = CT). Boldface indicates the group with the higher frequency. Asterisks are used to indicate the significance of the difference between the groups.

When working with clinical data, careful examination of the features can be beneficial. By comparing features with previous findings in the literature on agrammatism, we can be confident that we are measuring real effects and not just artifacts

	Rule	AG freq	CT freq	p
1	PP → IN NP	10.3	**24.9**	***
2	ROOT → NP	**2.9**	0.2	***
3	NP → DT NN POS	0.0	**0.7**	*
4	NP → PRP$ JJ NN	0.5	**0.7**	*
5	VP → TO VP	4.2	**7.5**	*
6	NP → NNP	5.9	**6.6**	
7	VP → VB PP	1.1	**2.9**	**
8	VP → VP CC VP	1.1	**3.1**	**
9	NP → DT NN NN	1.0	**2.7**	**
10	VP → VBD VP	0.1	**0.5**	*
11	WHADVP → WRB	0.5	**1.4**	*
12	FRAG → NP .	**0.7**	0.0	**
13	NP → JJ NN	**0.7**	0.0	**
14	SBAR → WHNP S	1.7	**3.1**	*
15	NP → NP SBAR	1.6	**2.5**	
16	S → NP VP	7.8	**16.1**	**
17	NP → PRP$ JJ NNS	0.0	**0.5**	*
18	NP → PRP$ NN NNS	0.0	**0.6**	*
19	SBAR → WHADVP S	0.4	**1.2**	*
20	VP → VBN PP	0.4	**2.0**	*

Table 2: Top 20 features ranked by GainRatio using the clean transcripts. ($^{*}p < 0.05$, $^{**}p < 0.005$, $^{***}p < 0.0005$).

of the parsing algorithm. This can also potentially provide an opportunity to observe features of agrammatic speech that have not been examined in manual analyses. We examine the top-ranked features in Table 2 in some detail, especially as they relate to previous work on agrammatism. In particular, the top features suggest some of the following features of agrammatic speech:

- Reduced number of prepositional phrases. This is suggested by feature 1, PP → IN NP. It is also reflected in features 7 and 20.
- Impairment in using verbs. We can see in feature 2 (ROOT → NP) that there is a greater number of utterances consisting of only a noun phrase. Feature 12 is also consistent with this pattern (FRAG → NP .). We also observe a reduced number of coordinated verb phrases (VP → VP CC VP).
- Omission of grammatical morphemes and function words. The agrammatic speakers use fewer possessives (NP → DT NN POS). Feature 9 indicates that the control participants more frequently produce compound nouns with a determiner (often ***the glass slipper*** or ***the fairy godmother***). Feature 4 also suggests some difficulty with determiners, as the agrammatic participants produce fewer nouns modified by a possessive pronoun and an adjective. Contrast this with feature 13, which shows agrammatic speech is more likely to contain noun phrases containing just an adjective and a noun. For example, in the control narratives we are more likely to see phrases such as *her godmother . . . waves* ***her magic wand***, while in the agrammatic narratives phrases like *Cinderella had* ***wicked stepmother*** are more common.
- Reduced number of embedded clauses and phrases. Evidence for this can be found in the reduced number of wh-adverb phrases (WHADVP → WRB), as well as features 14, 15, and 19.

	NB	**SVM**	**RF**
Narrative length	.62	.56	.64
Binary, no weights	**.87**	**.87**	.77
Binary, *tf-idf*	.87	**.90**	.85
Binary, InfoGain	.82	**.90**	.74
Binary, GainRatio	**.90**	.82	.79
Frequency, no weights	**.90**	.85	.85
Frequency, *tf-idf*	**.85**	.82	.77
Frequency, InfoGain	**.90**	**.90**	.82
Frequncy, GainRatio	.90	**.92**	.74
SC metrics, no weights	**.85**	.77	.82
SC metrics, InfoGain	**.85**	.77	.79
SC metrics, GainRatio	**.85**	.77	.82

Table 3: Average classification accuracy using the clean transcripts. The highest classification accuracy for each feature set is indicated with boldface.

The results of our classification experiment on the clean data are shown in Table 3. The results are similar for the binary and frequency features, with the best result of .92 achieved using an SVM classifier and frequency features, with GainRatio weights. The best results using parse features (.85–.92) are the same or slightly better than the best results using SC features (.85), and both feature sets perform above baseline.

5.2 Effect of non-narrative speech

In this section we perform two additional experiments, using the raw and auto-cleaned transcripts.

	Rule	AG freq.	CT freq.	p
1	NP → DT NN POS	0.0	**0.5**	*
2	PP → IN NP	12.2	**26.1**	***
3	SBAR → WHADVP S	0.4	**1.5**	*
4	VP → VBD	0.75	**1.1**	
5	VP → TO VP	4.3	**7.3**	*
6	S → CC PP NP VP .	0.04	**0.5**	*
7	NP → PRP$ JJ NNS	0.04	**0.5**	*
8	VP → AUX VP	3.7	**6.0**	
9	ROOT → FRAG	**4.5**	0.7	**
10	ADVP → RB	9.8	**12.3**	
11	NP → NNP	4.4	**6.2**	*
12	NP → DT NN	15.0	**24.1**	**
13	VP → VB PP	1.2	**2.8**	*
14	VP → VP CC VP	1.0	**2.9**	*
15	WHADVP → WRB	0.6	**1.5**	*
16	VP → VBN PP	0.4	**2.0**	*
17	INTJ → UH UH	**3.5**	0.3	*
18	VP → VBP NP	**0.5**	0.0	*
19	NP → NNP NNP	**1.5**	0.5	**
20	S → CC ADVP NP VP .	1.3	**2.3**	

Table 4: Top 20 features ranked by GainRatio using the raw transcripts. Bold feature numbers indicate rules which did not appear in Table 2. ($^{*}p < 0.05$, $^{**}p < 0.005$, $^{***}p < 0.0005$).

We discuss the differences between the selected features in each case, and the resulting classification accuracies.

Using the raw transcripts, we find that the ranking of features is markedly different than with the human-annotated transcripts (Table 4, bold feature numbers). Examining these production rules more closely, we observe some characteristics of agrammatic speech which were not detectable in the annotated transcripts:

- Increased number of dysfluencies. We observe a higher number of consecutive fillers (INTJ → UH UH) in the agrammatic data, as well as a higher number of consecutive proper nouns (NP → NNP NNP), usually two attempts at Cinderella's name. Feature 18 (VP → VBP NP) also appears to support this trend, although it is not immediately obvious. Most of the control participants tell the story in the past tense, and if they do use the present tense then the verbs are often in the third-person singular (*Cinderella **finds her fairy godmother***). Looking at the data, we found that feature 18 can indicate a verb agreement error, as in *he **attend the ball***. However, in almost twice as many cases it indicates use of the discourse markers *I mean* and *you know*, followed by a repaired or target noun phrase.
- Decreased connection between sentences. Feature 6 shows a canonical NP VP sentence, preceded by a coordinate conjunction and a prepositional phrase. Some examples of this from the control transcripts include, *And at the stroke of midnight . . .* and *And in the process . . . .* The conjunction creates a connection from one utterance to the next, and the prepositional phrase indicates the temporal relationship between events in the story, creating a sense of cohesion. See also the similar pattern in feature 20, representing sentence beginnings such as *And then . . . .*

However, there are some features which were highly ranked in the clean transcripts but do not appear in Table 4. What information are we losing by using the raw data? One issue with using the raw transcripts is that the inclusion of filled pauses "splits" the counts for some features. For example, the feature FRAG → NP . is ranked 12th using the clean transcripts but does not appear in the top 20 when using the raw transcripts. When we examine the transcripts, we find that the phrases that are counted in this feature in the clean transcripts are actually split into three features in the raw transcripts: FRAG → NP ., FRAG → INTJ NP ., and FRAG → NP INTJ ..

The classification results for the raw transcripts are given in Table 5. The results are similar to those for the clean transcripts, although in this case the best accuracy (.92) is achieved in three different configurations (all using the SVM classifier). The phrase-level features out-perform the traditional SC measures in only half the cases.

Using the auto-cleaned transcripts, we see some similarities with the previous cases (Table 6). However, some of the highly ranked features which disappeared when using the raw transcripts are now significant again (e.g. ROOT → NP, FRAG → NP .). There are also three remaining features which are significant and have not yet been discussed. Feature 9 shows an increased use of determiners with proper nouns (e.g. *the Cinderella*), a frank grammatical error. Feature 20

	NB	SVM	RF
Narrative length	.51	.62	.69
Binary, no weights	.87	**.92**	.82
Binary, *tf-idf*	.87	**.92**	.72
Binary, InfoGain	.85	**.87**	.82
Binary, GainRatio	.82	**.87**	.85
Frequency, no weights	.85	**.90**	.69
Frequency, *tf-idf*	.82	**.92**	.90
Frequency, InfoGain	**.85**	.74	**.85**
Frequncy, GainRatio	**.85**	.74	.82
SC metrics, no weights	.74	.79	**.82**
SC metrics, InfoGain	.77	**.85**	**.85**
SC metrics, GainRatio	.77	.85	**.87**

Table 5: Average classification accuracy using raw transcripts. The highest classification accuracy for each feature set is indicated with boldface.

provides another example of a sentence fragment with no verb. Finally, feature 19 represents an increased number of sentences or clauses consisting of a noun phrase followed by adjective phrase. Looking at the transcripts, this is not generally indicative of an error, but rather use of the word *okay*, as in *she dropped **her shoe okay***.

The classification results for the auto-cleaned data, shown in Table 7, show a somewhat different pattern from the previous experiments. The accuracies using the parse features are generally higher, and the best result of .97 is achieved using the binary features and the naive Bayes classifier. Interestingly, this data set also results in the lowest accuracy for the syntactic complexity metrics.

5.3 Phrase-level parse features

The classifiers in Tables 3, 5, and 7 used the phrase-level parse features as well as the CFG productions. Although these features were calculated for NPs, VPs, and PPs, the NP features were never selected by the GainRatio ranking algorithm, and did not differ significantly between groups. The significance levels of the VP and PP features are reported in Table 8. PP rate and proportion are significantly different in all three sets of transcripts, which is consistent with the high ranking of PP → IN NP in each case. VP rate and proportion are often significant, although less so. Notably, PP and VP length are both significant in the clean transcripts, but not significant in the raw transcripts and only barely significant in the auto-cleaned transcripts.

	Rule	AG freq.	CT freq.	p
1	PP → IN NP	12.0	**26.0**	***
2	NP → DT NN POS	0.0	**0.7**	*
3	VP → VP CC VP	0.8	**2.9**	**
4	S → CC SBAR NP VP .	0.0	**0.5**	
5	SBAR → WHADVP S	0.4	**1.5**	*
6	NP → NNP	5.6	**6.7**	
7	VP → VBD	0.8	**1.1**	
8	S → CC PP NP VP .	0.04	**0.6**	*
9	NP → DT NNP	**0.6**	0.0	**
10	VP → TO VP	4.6	**7.5**	*
11	ROOT → FRAG	**3.0**	0.5	***
12	ROOT → NP	**2.1**	0.1	*
13	VP → VBP NP	1.7	**3.6**	
14	NP → PRP$ JJ NNS	0.04	**0.5**	*
15	VP → VB PP	1.1	**2.8**	**
16	VP → VBN PP	0.4	**1.9**	*
17	FRAG → NP .	**0.4**	0.0	*
18	NP → NNP .	**2.1**	0.1	
19	S → NP ADJP	**0.4**	0.0	*
20	FRAG → CC NP .	**0.7**	0.07	**

Table 6: Top 10 features ranked by GainRatio using the auto-cleaned transcripts. Bold feature numbers indicate rules which did not appear in Table 2. ($^{*}p < 0.05$, $^{**}p < 0.005$, $^{***}p < 0.0005$).

5.4 Analysis of variance

With a multi-way ANOVA we found significant main effects of classifier ($F(2,63) = 11.6$, $p < 0.001$) and data set ($F(2,63) = 11.2$, $p < 0.001$) on accuracy. A Tukey post-hoc test revealed significant differences between SVM and RF ($p < 0.001$) and NB and RF ($p < 0.001$) but not between SVM and NB. As well, we see a significant difference between the clean and auto-cleaned data ($p < 0.001$) and the raw and auto-cleaned data ($p < 0.001$) but not between the raw and clean data. There was no significant main effect of weighting scheme or feature type (binary or frequency) on accuracy. We did not examine any possible interactions between these variables.

6 Discussion

6.1 Transcripts

We achieved the highest classification accuracies using the auto-cleaned transcripts. The raw transcripts, while containing more information about dysfluent events, also seemed to cause more dif-

	NB	SVM	RF
Narrative length	.51	.62	.64
Binary, no weights	.92	**.95**	.90
Binary, *tf-idf*	.92	**.95**	.87
Binary, InfoGain	**.97**	.90	.85
Binary, GainRatio	**.97**	.90	.95
Frequency, no weights	.90	**.95**	.77
Frequency, *tf-idf*	.87	**.95**	.79
Frequency, InfoGain	**.92**	.85	.82
Frequncy, GainRatio	.92	.87	**.95**
SC metrics, no weights	**.79**	.77	.74
SC metrics, InfoGain	**.79**	.74	.72
SC metrics, GainRatio	**.79**	.74	.67

Table 7: Average classification accuracy using auto-cleaned transcripts. The highest classification accuracy for each feature set is indicated with boldface.

	Clean	Raw	Auto
PP rate	***	***	***
PP proportion	***	***	**
PP length	**		
VP rate		**	*
VP proportion	***	*	*
VP length	***		*

Table 8: Significance of the phrase-level features in each of the three data sets ($^{*}p < 0.05$, $^{**}p < 0.005$, $^{***}p < 0.0005$).

ficulty for the parser, which mis-labelled filled pauses and false starts in some cases. We also found that the insertion of filled pauses resulted in the creation of multiple features for a single underlying grammatical structure. The auto-cleaned transcripts appeared to avoid some of those problems, while still retaining information about many of the non-narrative speech productions that were removed from the clean transcripts.

Some of the features from the auto-cleaned transcripts appear to be associated with the discourse level of language, such as connectives and discourse markers. A researcher solely interested in studying the syntax of language might resist the inclusion of such features, and prefer to use only features from the human-annotated clean transcripts. However, we feel that such productions are part of the grammar of spoken language, and merit inclusion. From a practical standpoint, our findings are reassuring: data preparation that can be done automatically is much more feasible in many situations than human annotation.

6.2 Features

CFG production rules can offer a more detailed look at specific language impairments. We were able to observe a number of important characteristics of agrammatic language as reported in previous studies: fragmented speech with a higher incidence of solitary noun phrases, difficulty with determiners and possessives, reduced number of prepositional phrases and embedded clauses, and (in the raw transcripts), increased use of filled pauses and repair phrases. For this reason, we believe that they are more useful for the analysis of disordered or otherwise atypical language than traditional measures of syntactic complexity.

In some cases an in-depth analysis may not be required, and in such cases it may be tempting to simply use one of the more-general syntactic complexity measures. Nevertheless, even in our simple binary classification task, we found that using the more-specific features gave us a higher accuracy.

6.3 Future work

Because of the limited data, we consider these results to be preliminary. We hope to replicate this study as more data become available in the future. We also plan to examine the effect, if any, of the specific narrative task. Furthermore, we have shown that these methods are effective for the analysis of agrammatic aphasia, but there are other types of aphasia in which semantic, rather than syntactic, processing is the primary impairment. We would like to extend this work to find features which distinguish between different types of aphasia.

Although we included manually transcribed data in this study, these methods will be most useful if they are also effective on automatically recognized speech. Previous work on speech recognition for aphasic speech reported high error rates (Fraser et al., 2013a). Our finding that the auto-cleaned transcripts led to the highest classification accuracy is encouraging, but we will have to test the robustness to recognition errors and the dependence on sentence boundary annotations.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada and National Institutes of Health R01DC01948 and R01DC008552.

References

Harald Baayen, Hans Van Halteren, and Fiona Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132.

Roelien Bastiaanse and Cynthia K. Thompson. 2012. *Perspectives on Agrammatism*. Psychology Press.

Julian Brooke and Graeme Hirst. 2012. Robust, lexicalized native language identification. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 391–408.

Jieun Chae and Ani Nenkova. 2009. Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 139–147.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, pages 132–139.

Franca Debole and Fabrizio Sebastiani. 2004. Supervised term weighting for automated text categorization. In *Text mining and its applications*, pages 81–97. Springer.

Kathleen Fraser, Frank Rudzicz, Naida Graham, and Elizabeth Rochon. 2013a. 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. 2013b. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*.

Harold Goodglass, Julie Ann Christiansen, and Roberta E. Gallagher. 1994. Syntatic constructions used by agrammatic speakers: Comparison with conduction aphasics and normals. *Neuropsychology*, 8(4):598.

Khairun-nisa Hassanali, Yang Liu, Aquiles Iglesias, Thamar Solorio, and Christine Dollaghan. 2013. Automatic generation of the index of productive syntax for child language transcripts. *Behavior research methods*, pages 1–9.

David I. Holmes and Sameer Singh. 1996. A stylometric analysis of conversational speech of aphasic patients. *Literary and Linguistic Computing*, 11(3):133–140.

Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. 2011. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.

Ljiljana Progovac. 2006. *The Syntax of Nonsententials: Multidisciplinary Perspectives*, volume 93. John Benjamins.

Emily T. Prud'hommeaux, Brian Roark, Lois M. Black, and Jan van Santen. 2011. Classification of atypical language in autism. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, CMCL '11, pages 88–96.

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.

Elizabeth Rochon, Eleanor M. Saffran, Rita Sloan Berndt, and Myrna F. Schwartz. 2000. Quantitative analysis of aphasic sentence production: Further development and new data. *Brain and Language*, 72(3):193–218.

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.

Kenji Sagae, Alon Lavie, and Brian MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 197–204.

Ben Swanson and Eugene Charniak. 2012. Native language detection with tree substitution grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 193–197.

Cynthia K. Thompson, Lewis P. Shapiro, Ligang Li, and Lee Schendel. 1995. Analysis of verbs and verb-argument structure: A method for quantification of aphasic language production. *Clinical Aphasiology*, 23:121–140.

Sze-Meng Jojo Wong and Mark Dras. 2010. Parser features for sentence grammaticality classification. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 67–75.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610.

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