CORTEX XXX (2013) 1–18

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Special issue: Research report

Automated classification of primary progressive aphasia subtypes from narrative speech transcripts

Kathleen C. Fraser [a,*], *Jed A. Meltzer* [b], *Naida L. Graham* [c,d], *Carol Leonard* [e], *Graeme Hirst* [a], *Sandra E. Black* [f,g] *and Elizabeth Rochon* [c,d]

[a] *Department of Computer Science, University of Toronto, Toronto, Ontario, Canada*
[b] *Rotman Research Institute, Baycrest Centre, Toronto, Ontario, Canada*
[c] *Department of Speech-Language Pathology, University of Toronto, Toronto, Ontario, Canada*
[d] *Toronto Rehabilitation Institute, Toronto, Ontario, Canada*
[e] *School of Rehabilitation Sciences, University of Ottawa, Ottawa, Ontario, Canada*
[f] *L.C. Campbell Cognitive Neurology Research Unit, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada*
[g] *Department of Medicine (Neurology), University of Toronto, Toronto, Ontario, Canada*

ARTICLE INFO

Article history:
Received 29 June 2012
Reviewed 13 September 2012
Revised 13 November 2012
Accepted 6 December 2012
Published online xxx

Keywords:
Semantic dementia
Progressive nonfluent aphasia
Narrative speech
Natural language processing
Machine learning

ABSTRACT

In the early stages of neurodegenerative disorders, individuals may exhibit a decline in language abilities that is difficult to quantify with standardized tests. Careful analysis of connected speech can provide valuable information about a patient's language capacities. To date, this type of analysis has been limited by its time-consuming nature. In this study, we present a method for evaluating and classifying connected speech in primary progressive aphasia using computational techniques. Syntactic and semantic features were automatically extracted from transcriptions of narrative speech for three groups: semantic dementia (SD), progressive nonfluent aphasia (PNFA), and healthy controls. Features that varied significantly between the groups were used to train machine learning classifiers, which were then tested on held-out data. We achieved accuracies well above baseline on the three binary classification tasks. An analysis of the influential features showed that in contrast with controls, both patient groups tended to use words which were higher in frequency (especially nouns for SD, and verbs for PNFA). The SD patients also tended to use words (especially nouns) that were higher in familiarity, and they produced fewer nouns, but more demonstratives and adverbs, than controls. The speech of the PNFA group tended to be slower and incorporate shorter words than controls. The patient groups were distinguished from each other by the SD patients' relatively increased use of words which are high in frequency and/or familiarity.

* *Corresponding author*. Department of Computer Science, University of Toronto, 10 King's College Road, Room 3302, Toronto, Ontario, Canada M5S 3A6.
E-mail address: kfraser@cs.toronto.edu (K.C. Fraser).

0010-9452/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cortex.2012.12.006

1. Introduction

Primary progressive aphasia (PPA) is a dementia syndrome, resulting from neurodegenerative disease, in which language impairment is the earliest and most salient feature. It is widely accepted that there are three variants of PPA (Gorno-Tempini et al., 2004): progressive nonfluent aphasia (PNFA), progressive fluent aphasia, often referred to as semantic dementia (SD) due to the pervasive semantic impairment, and logopenic progressive aphasia. PNFA is characterized by nonfluent, hesitant, effortful speech, with word-finding difficulty; in addition, agrammatism and/or apraxia of speech are considered to be core features. In SD, there is severe anomia, although spoken output remains fluent, well-articulated, and grammatically correct, with normal prosody. The logopenic variant is associated with hesitant speech, obvious word-finding difficulty, and intact word repetition but poor repetition of phrases and sentences; this variant is not a focus of the present study and therefore will not receive further attention. Until recently, most systematic investigations of spoken output in PPA focused on single word production (naming, reading, repetition), but there is now a small literature that examines production of connected speech. Difficulty with conversing is often a presenting complaint in PPA, and diagnostic criteria describe the nature of the impairment in spoken output that is indicative of each variant (Gorno-Tempini et al., 2004). Because impairment in connected speech is the essence of PPA, thorough characterization seems essential. The main hurdle to date has been the laborious process required for transcription and systematic analysis of connected speech. Nevertheless, progress has been made and we are beginning to understand the characteristics of language production in connected speech in each variant of PPA.

Patients with PNFA tend to have reduced output in comparison with control participants: it has been shown that they produce fewer words (Graham et al., 2004; Wilson et al., 2010), shorter phrase length (Knibb et al., 2009), and a shorter mean length of utterance (Ash et al., 2006; Thompson et al., 2012). As well, their speech rate is slower and their speech is less informative than that of controls (Ash et al., 2006; Graham et al., 2004; Knibb et al., 2009; Thompson et al., 2012; Wilson et al., 2010). Impairment in grammatical competency is an established feature of the syndrome. These patients produce increased grammatical errors (Knibb et al., 2009), fewer grammatically correct sentences (Thompson et al., 2012), and show impaired production of verb inflection and argument structure (Thompson et al., 2012). The degree of grammatical impairment is a matter for debate, as not all patients show agrammatism and production of normal proportions of content and function words has been documented (Graham et al., 2004). Knibb et al. (2009) noted that increased grammatical errors and simplified syntax were universal in the PNFA patients they studied, while pervasive agrammatism was not common.

The work on production of connected speech in patients with SD has demonstrated that they tend to use words which are higher in frequency but less specific than the words used by controls (Meteyard and Patterson, 2009). They also produce more pronouns, as well as more pronouns with ambiguous referents (Kavé et al., 2007; Meteyard and Patterson, 2009; Patterson and MacDonald, 2006; Wilson et al., 2010). Thus, it is not surprising that the speech of SD patients has been shown to be less informative than that of controls (Ash et al., 2006; Kavé et al., 2007; Meteyard and Patterson, 2009). There is also a tendency to use nouns and verbs which are higher in frequency than those used by controls (Bird et al., 2000). The rate of syntactic and phonological errors is no higher than controls (Sajjadi et al., 2012; Wilson et al., 2010), but the level of syntactic ability remains unclear. Some studies have documented normal ratios of content words to function words and of nouns to verbs (Meteyard and Patterson, 2009; Sajjadi et al., 2012), suggesting normal grammatical production, but others found that both of these ratios were abnormal (Bird et al., 2000; Garrard and Forsyth, 2010; Thompson et al., 2012). Similarly, there has been inconsistency with respect to the findings regarding speech rate, which has been found to be both normal (Bird et al., 2000; Garrard and Forsyth, 2010; Meteyard and Patterson, 2009; Thompson et al., 2012) and reduced (Ash et al., 2006; Sajjadi et al., 2012; Wilson et al., 2010). Interestingly, Sajjadi et al. (2012) found that SD patients do not exhibit frequent circumlocution, despite numerous clinical descriptions to the contrary.

In this study, we examine narrative speech in PNFA and SD. In contrast to the studies reviewed above, to gain maximum information we used methods from natural language processing, which involves the use of software to analyze speech samples, or in our case, transcriptions of speech samples. These methods enable, for example, part-of-speech (POS) tags to be automatically assigned to words in a text using a statistical POS tagger. Others have begun to use these methods to analyze spoken output in dementia. For example, Roark et al. (2011) compared automatic and manual methods for determining syntactic structure of spoken output, and demonstrated that the automatic method was sufficiently accurate to enable identification of syntactic complexity measures that distinguished between healthy participants and those with mild cognitive impairment.

Peintner et al. (2008) have adopted this approach. They studied speech from patients with frontotemporal dementia (FTD), and used a subset of extracted features as input to machine learning classifiers to classify each participant as belonging to the PNFA, SD, or behavioural variant FTD groups, or as a control. A similar procedure was followed by Jarrold et al. (2010) when they used machine learning algorithms to classify transcriptions of speech from participants with pre-symptomatic Alzheimer's disease (AD), mild cognitive impairment, or depression. Both studies had some success with classification based on samples of connected speech, but they are limited in that they do not report which features were able to reliably distinguish between patient groups.

The present study had two aims. The first was to develop a machine learning classifier that would analyze speech samples and be able to distinguish between control participants and participants with PNFA or SD, as well as between the two patient groups. The other aim of this study was to identify the automatically extracted features that best distinguish the groups, and to compare this with results in the

Please cite this article in press as: Fraser KC, et al., Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2012.12.006

literature that are based on traditional (manual) analysis methods. Identification of the distinguishing features is important for improved detection and differentiation of the variants of PPA.

2. Participants and methods

2.1. Participants

Our participants comprised 24 patients diagnosed with either the fluent (SD) or nonfluent variant (PNFA) of PPA, and 16 age- and education-matched healthy controls. The patient group is an unselected sample of patients with SD or PNFA, except that participants who were unable to complete the narrative task ($n = 7$) were excluded: 2 PNFA patients had incomprehensible speech, 1 PNFA patient said nothing, 1 SD patient refused to attempt the task, and the responses of 1 PNFA and 2 SD patients did not include any of the story that they were asked to tell, but instead comprised statements of how they could not remember the story. Participants with PPA were recruited through three memory clinics in Toronto and each was diagnosed by an experienced behavioural neurologist. There were a further 7 patients in the cohort who were diagnosed with logopenic PPA, but this group was not included in the present study due to its small size. Control participants were recruited from a volunteer participant pool. All participants were native speakers of English, or completed some of their education in English. Exclusion criteria included a known history of drug or alcohol abuse, or a history of neurological or major psychiatric illness.

The study was approved by the Research Ethics Boards of all the hospitals involved in recruitment, as well as the board at the University of Toronto. Written informed consent was obtained from all participants.

Diagnosis was based on history, neuroimaging, neurological examination and neuropsychological testing, and all patients met current criteria for PPA (Gorno-Tempini et al., 2011). Patients with the fluent variant exhibited grammatically correct fluent speech, with word-finding difficulties. Those with the nonfluent variant had effortful, halting speech with anomia, although not all exhibited clear agrammatism in production or clear apraxia of speech on formal testing. Demographic data are listed in Table 1.

All participants underwent a battery of neuropsychological and linguistic tests as part of a longitudinal study of PPA which is being conducted in the Department of Speech-Language Pathology of the University of Toronto. The neuropsychological test information is reported in Table 1. The level of general cognitive functioning was measured using the Mini-Mental State Examination (Folstein et al., 1975) and the Dementia Rating Scale-R (Jurica et al., 2001). The two patient groups did not differ on these tests, but were both impaired relative to controls. In keeping with the diagnosis of PPA, both patient groups performed poorly on a test of picture naming (Boston Naming Test, Kaplan et al., 2001) and on category fluency for animals (where participants are asked to name all the animals they can think of in 1 min). Impairment in syntactic comprehension is a known feature of PNFA and indeed was exhibited by the PNFA group studied here; this ability was measured using the Test for the Reception of Grammar (Bishop, 2003), and the nonfluent group (only) performed significantly worse than controls. Impairment in single word comprehension is an established feature of SD; both of our patient groups were impaired on single word comprehension, but as expected, the SD patients performed

Table 1 – Demographic and neuropsychological data for each participant group. Values shown are mean (standard deviation). Asterisks denote significant effect of group on 1-way analyses of variance at $^{*}p < .05$, $^{***}p < .001$.

	SD ($n = 10$)	PNFA ($n = 14$)	Controls ($n = 16$)	Group effect
Demographic information				
Age	65.6 (7.4)	64.9 (10.1)	67.8 (8.2)	
Years of education	17.5 (6.1)	14.3 (3.6)	16.8 (4.3)	
Sex	3 F	6 F	7 F	
Handedness	9 R	13 R	16 R	
General cognitive function				
Mini-Mental State Examination (/30)	24.4 (4.3)[a]	25.0 (2.9)[a]	29.3 (.8)	***
Dementia Rating Scale-R (/144)	117.2 (12.6)[a]	123.9 (15.6)[a]	142.2 (1.7)	***
Language production				
Boston Naming (/60)	13.9 (7.3)[a,b]	39.6 (11.5)[a]	55.8 (3.3)	***
Category fluency – animals	7.6 (3.7)[a]	12.3 (6.0)[a]	20.4 (4.4)	***
Language comprehension				
Test for the Reception of Grammar (/80)	71.4 (11.0)	63.9 (12.0)[a]	79.1 (.9)	***
Peabody Picture Vocabulary Test (/204)	113.8 (30.8)[a,b]	172.9 (14.3)[a]	196.1 (3.9)	***
Visuospatial				
Copy of Rey Complex Figure (/36)	33.2 (2.6)	29.9 (5.3)[a]	33.4 (1.4)	*
VOSP cube analysis subtest (/10)	9.4 (1.9)	9.2 (1.6)	8.5 (2.1)	
Nonverbal memory				
30 min recall of Rey Complex Figure (/36)	12.7 (7.1)	14.9 (6.3)	18.8 (6.9)	
Nonverbal reasoning				
Raven's Coloured Progressive Matrices (/36)	31.5 (5.0)	27.1 (6.5)[a]	31.8 (4.2)	*

a Significantly different from controls.
b Significantly different from nonfluent patients.

Please cite this article in press as: Fraser KC, et al., Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2012.12.006

significantly worse than the PNFA patients (Peabody Picture Vocabulary Test, Dunn and Dunn, 1997). Consistent with the diagnosis of PPA, performance was generally better on nonverbal tests. The PNFA group was mildly impaired on copying the Rey Complex Figure (Rey, 1941) while the SD group showed normal performance. On another measure of visuospatial functioning, the cube analysis subtest from the Visual Object and Space Perception Battery (Warrington and James, 1991), both patient groups performed normally. Similarly, performance on nonverbal episodic memory was normal for both patient groups; this was assessed by asking participants to recall the Rey Complex Figure 30 min after copying it. Finally, nonverbal reasoning was relatively preserved, although it was mildly impaired for the PNFA group (only) (Raven's Coloured Progressive Matrices, Raven, 1962).

2.2. Narrative task

Speech samples were elicited by having participants tell the Cinderella story, as in Saffran et al. (1989). To prompt their memories for the story, participants were given as much time as they needed to examine a picture book illustrating the story. When each participant had finished looking at the pictures, and the book had been removed, the examiner said "Now you tell me the story. Include as much detail as you can and try to use complete sentences." After letting the participant speak for as long as he or she wished, if the story was incomplete, general encouragement for more speech was given, for example, "Good, tell me more about that", "What happens next", "Go on", etc. At no time were specific questions or prompts given. The narratives were recorded on a digital audio recorder for subsequent verbatim transcription. Transcription was done in accordance with the procedures used in the Quantitative Production Analysis (Berndt et al., 2000), with the exception that punctuation and sentence initial capitalization were used. Brief pauses were marked with commas, while pauses longer than 1 sec were timed and the length of the pause was noted [e.g., (2 sec)]; however, commas and pauses were removed before analysis. Sentence boundaries were marked with full stops. Placement of sentence boundaries was guided by semantic, syntactic and prosodic features, using a method essentially identical to that described by Thompson et al. (2012). When utterance boundaries were ambiguous, we created shorter utterances [as did Thompson et al. (2012) and Wilson et al. (2010)]. Fillers such as *um* and *uh* were transcribed (and analyzed), but were not included in the total word count. Repetitions, false starts and repeated but incomplete attempts at a given word were transcribed, but only repetitions of words and false starts were included in the total word count. Neologisms were transcribed with the International Phonetic Alphabet, and words/passages which were incomprehensible were marked with [###]. Phonemic errors were written using the Roman alphabet and were followed by the transcriber's gloss of the word, which was put into double brackets. Neologisms and incomprehensible speech were not included in the automated analyses or in the word counts as we could not be certain how many words were represented; note, however, that incomprehensible passages were always brief. There were only rare instances of neologisms or incomprehensible speech: 3 participants (one in each group) had 2 occurrences each, while a further 3 participants had 1 occurrence (2 PNFA, 1 SD).

2.3. Analysis of narrative speech

The automatically extracted features are defined in Table 2. The first feature is the number of words in the transcript. The subsequent 22 structural features (2–23) were calculated using Lu's L2 Syntactic Complexity Analyzer (Lu, 2010), which uses the Stanford parser (Klein and Manning, 2003). We modified features 3–9 to be normalized by the total number of words, to facilitate comparison between narratives of different lengths. Lu used these features to analyze the syntactic complexity of college-level English essays from Chinese students, but the software has also been used to analyze spoken language (Chen and Zechner, 2011). We have not attempted to adapt this tool specifically for the study of aphasic speech; rather, we are interested to see how well such methods perform in the absence of any domain-adaptation.

The next four features (24–27) are also measures of syntactic complexity. Tree-based measures have been used to detect age-related cognitive decline (Cheung and Kemper, 1992) and mild cognitive impairment (Roark et al., 2011). Parse trees were constructed using the Stanford parser, so they are based on the same structural model as features 2–23. The Yngve depth quantifies to what extent the syntactic structure of a sentence contains left-branching rather than right-branching phrases, which provides a comparison metric of syntactic complexity (Yngve, 1960). For detailed illustrations of how Yngve depth is quantified, see Sampson (1997), Cheung and Kemper (1992), or Yngve (1960). We quantified Yngve depth as mean depth over all words, the maximum depth in the sentence, and the total depth. For example, a sentence with an object-embedded relative clause such as *The juice that the child spilled stained the rug* is more left-branching than one with a subject-embedded relative clause such as *The child spilled the juice that stained the rug* (Stromswold et al., 1996). Using our procedures, the first sentence is assigned these values: (max depth: 3, mean depth: 1.67, total depth: 15), while the second sentence is assigned: (max depth: 2, mean depth: 1.11, total depth: 10).

Features 28–40 rely explicitly on the POS tags for each word in the sample, determined by using the Stanford POS tagger (Toutanova et al., 2003). Differences in the noun and verb production of SD and PNFA patients have been noted before (Harciarek and Kertesz, 2011). It has also been observed that PNFA patients are more likely to omit inflections and function words (Harciarek and Kertesz, 2011). Here, function words included determiners, pronouns, prepositions, conjunctions, particles, and modals.

Word-level phonemic errors can present a potential problem to the tagger. We have introduced an extra tag called *not in dictionary* or *NID* for cases in which the speaker produces a nonword token. This prevents such tokens from being counted towards the wrong POS category. If a word error results in another English word, this is not detected by our system and could be tagged incorrectly. However, the context around the word may provide useful clues to the tagger in such cases.

Please cite this article in press as: Fraser KC, et al., Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2012.12.006

Table 2 – Definitions of features.

Feature		Definition
1	Words	
2	Sentences	
3	T-units	A clause and all of its dependent clauses
4	Clauses	A structure consisting of at least a subject and a finite verb
5	Coordinate phrases	A phrase immediately before a coordinating conjunction
6	Complex nominals	A noun phrase, clause, or gerund that stands in for a noun
7	Complex T-units	A T-unit which contains a dependent clause
8	Verb phrases	A phrase consisting of at least a verb and its dependents
9	Dependent clauses	A clause which could not form a sentence on its own
10	Mean length of sentence	
11	Mean length of clause	
12	Mean length of T-unit	
13	Dependent clauses per clause	
14	Dependent clauses per T-unit	
15	Verb phrases per T-unit	
16	Clauses per sentence	
17	Clauses per T-unit	
18	Complex T-units per T-unit	
19	Coordinate phrases per T-unit	
20	Complex nominals per T-unit	
21	T-units per sentence	
22	Coordinate phrases per clause	
23	Complex nominals per clause	
24	Tree height	Height of the parse tree
25	Total depth	Total Yngve depth
26	Max depth	Maximum Yngve depth
27	Mean depth	Mean Yngve depth
28	Nouns	# Nouns/# words
29	Verbs	# Verbs/# words
30	Noun–verb ratio	# Nouns/# verbs
31	Noun ratio	# Nouns/(# nouns + # verbs)
32	Inflected verbs	# Inflected verbs/# verbs
33	Light verbs	# Light verbs/# verbs
34	Determiners	# Determiners/# words
35	Demonstratives	# Demonstratives/# words
36	Prepositions	# Prepositions/# words
37	Adjectives	# Adjectives/# words
38	Adverbs	# Adverbs/# words
39	Pronoun ratio	# Pronouns/(# nouns + # pronouns)
40	Function words	# Function words/# words
41	Frequency	Mean frequency of all words appearing in the frequency norms
42	Noun frequency	Mean frequency of nouns appearing in the frequency norms
43	Verb frequency	Mean frequency of verbs appearing in the frequency norms
44	Imageability	Mean imageability of all words appearing in the imageability norms
45	Noun imageability	Mean imageability of nouns appearing in the imageability norms
46	Verb imageability	Mean imageability of verbs appearing in the imageability norms
47	Age of acquisition	Mean age of acquisition of all words appearing in the age of acquisition norms
48	Noun age of acquisition	Mean age of acquisition of nouns appearing in the age of acquisition norms
49	Verb age of acquisition	Mean age of acquisition of verbs appearing in the age of acquisition norms
50	Familiarity	Mean familiarity of all words appearing in the familiarity norms
51	Noun familiarity	Mean familiarity of nouns appearing in the familiarity norms
52	Verb familiarity	Mean familiarity of verbs appearing in the familiarity norms
53	Type-token ratio	# Unique word types/# words
54	Word length	Mean number of letters in each word
55	Fillers	# Fillers/# words
56	Um	# Occurrences of 'um'/# words
57	Uh	# Occurrences of 'uh'/# words
58	Speech rate	# Words uttered/total time in minutes

Verbs can be categorized as being *heavy* or *light*, according to their semantic complexity. Light verbs like *have* or *do* can be used in such a wide variety of different contexts that they are similar in some ways to closed-class function words (Breedin et al., 1998). We used the same list of light verbs as Breedin et al. (1998), namely: *be*, *have*, *come*, *go*, *give*, *take*, *make*, *do*, *get*, *move*, and *put*. All verbs which are not on this list are considered to be heavy.

Features 41–52 measure frequency, imageability, age of acquisition, and familiarity. Frequency was calculated according to the SUBTL norms (Brysbaert and New, 2009), and the remaining three according to the combined Bristol norms and Gilhooly–Logie norms (Stadthagen-Gonzalez and Davis, 2006; Gilhooly and Logie, 1980). In addition to calculating the overall averages, the measures were calculated for nouns and verbs independently, to explore any possible dissociations. We calculated the proportion of words covered by the norms based on unique word forms (as opposed to individual occurrences). The coverage for the frequency norms is excellent – between .92 and .95 across the three groups. The coverage for the imageability, age of acquisition, and familiarity norms is not as good, ranging from .25 to .31 for all content words across the three groups. One reason for this is that the Bristol norms were specifically designed to exclude high frequency words, as the authors wanted to use words in "the frequency range most often sampled by psycholinguistic experiments" (Stadthagen-Gonzalez and Davis, 2006). This means that most of the words included in the norms have

Please cite this article in press as: Fraser KC, et al., Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2012.12.006

frequencies between 1 and 100 counts per million. For example, nouns like *thing* (which has a frequency of 1088 counts per million in the SUBTL norms) and *name* (641 counts per million) are excluded, as are verbs like *go* (3793 counts per million) and *do* (6135 counts per million).

The last six features (53–58) are measures of fluency and vocabulary richness. One way to measure vocabulary size is by calculating the type-token ratio, which is the ratio of the number of word types to the total number of words in the sample. A type-token ratio of 1.0 would mean that every word in the sample was unique; a low type-token ratio would indicate that many words were repeated. Filled pauses are measured by counting occurrences of the words *um*, *uh*, *ah*, and *er*, called "fillers" in Table 2. The words *um* and *uh* were also counted individually, because of research which suggests that they may be used to indicate major and minor pauses, respectively (Clark and Fox Tree, 2002). Finally, speech rate has been shown to distinguish between PPA patients and controls, although the results in SD are inconsistent (see Introduction). Here we consider an estimate of the speech rate, which was calculated by dividing the number of words produced by the participant by the total speech sample time.

Impaired speech can present difficulties for automatic language processing techniques, which have typically been developed for well-formed, written text. However, these methods represent a starting point for future development of more sophisticated techniques. To increase the probability of the structural analyses producing accurate results, our system counts and then removes the filled pauses from the transcript. Short nonword tokens (i.e., repeated but incomplete attempts at a given word, e.g., *br- bring*) are also removed.

To test the results of our automatic methods against traditional manual methods, we had a human annotator perform POS tagging and calculate a subset of the parse measures for three randomly-chosen narratives from each of the three participant groups (22.5% of the total data set). For POS tagging, we measured the agreement between the human annotator and the automatic tagger by counting the number of tags on which they both agreed, and dividing by the total number of tags. The average agreement was 87.3% for the PNFA group, 89.2% for the SD group, and 91.9% for the control group. For comparison, the best reported accuracies for statistical taggers on written, well-formed text are around 97% (Manning, 2011), while Pakhomov et al. (2010) reported a tagging accuracy of 86% on PNFA speech transcripts.

In comparing the parse features, we were limited by the fact that Lu's program simply outputs counts for each measure, rather than the actual constituents being measured. This makes it impractical to use traditional parse measures such as the PARSEVAL measures (Manning and Schütze, 1999). Instead we had the annotator produce counts of different structures, just as the software does, and then correlated the two sets of counts. We examined a set of features that seemed to be important in distinguishing the groups based on preliminary analyses, although not all of them were significant in the final version of the system. We measured the correlation in two ways: the correlation between the counts for each individual sentence, and the correlation between the total counts for each narrative. The correlation coefficients for these measures are shown in Table 3.

Correlations between the scores calculated by the human annotator and Lu's system were high when considered across narratives, which is the most relevant comparison for our purposes, as only the total scores for the narratives were used as input to the classifiers. The per-sentence correlations were somewhat lower. Inspection of the discrepancies between the manual and automatic scoring revealed a systematic pattern. In general, the automatic system has high agreement with the human annotator when determining the number of clauses. However, it has difficulty labelling clauses as being either independent or dependent, especially when the clauses are not connected by a conjunction. In many such cases, we found that the system counted a different number of dependent clauses than the human annotator, which in turn affected the number of T-units and complex T-units. Given this apparently systematic error, the results from Lu's syntactic complexity analyzer must be interpreted with caution.

2.4. Classification

We trained machine learning classifiers to predict participants' diagnoses based on a set of features extracted automatically from speech transcripts. Including too many features risks overfitting the classifier to idiosyncrasies in the training set, resulting in poor generalization to new data points. Therefore, some process of feature selection is necessary. To select features on which to train the classifiers, we conducted a two-sample *t*-test on each feature between the two groups that were to be distinguished. All features that were significant at $p < .05$ were used for classification. The values of the selected features make up a feature vector, which defines a point in feature space. The goal of a machine learning classifier is to take a feature vector as input, and output a class label (in this case, either SD, PNFA, or control). Three machine learning classifiers from the WEKA machine learning toolkit were compared (Hall et al., 2009).

Naïve Bayes is a classifier based on Bayes's theorem. It is called "naïve" because it makes the strong simplifying assumption that all of the features are conditionally independent given the class. The classifier learns estimates for the class-conditional probabilities and priors for each class from the training data. In the classification stage, it uses Bayes's theorem to assign a data point to the class that maximizes the posterior probability. Naïve Bayes is widely used, even in cases where the independence assumption is known to be false, and often performs well. The rationale for this is that even though

Table 3 – Correlations between human- and computer-generated counts for syntactic structures.

	Clauses	Dependent clauses	T-units	Complex T-units	Coordinate phrases
Per narrative	.9966	.9319	.9269	.8792	.9475
Per sentence	.9630	.6396	.4756	.4568	.7921

Please cite this article in press as: Fraser KC, et al., Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2012.12.006

the probability estimates may be inaccurate, the classification results (which depend only on which probability is the highest, and not on the actual numbers) can still be good (Manning et al., 2008).

In contrast to naïve Bayes, which attempts to model the classes themselves, logistic regression is a discriminative classifier which attempts to model the boundary between the classes instead. Logistic regression estimates the posterior probability directly from the training data. Research suggests that naïve Bayes may perform better in cases where there is not a large amount of training data (Ng and Jordan, 2002). However, the benefit of logistic regression is that it does not assume the features are conditionally independent. Peintner et al. (2008) used logistic regression, along with two other classifiers not considered here, on various classification tasks involving FTD subtypes and healthy controls. They had mixed results, with logistic regression achieving the best results in two out of six cases.

Support vector machines (SVMs) are another type of linear discriminative classifier which have become very popular in natural language processing applications in the past several years (Manning et al., 2008). SVMs are maximum margin classifiers, which means they find the decision boundary between two classes that maximizes the margin between the two classes. In other words, they maximize the distance between the decision boundary and the nearest data points. If the data are not linearly separable, then the algorithm tries to maximize the margin while also minimizing the misclassification error (Manning et al., 2008).

The classifiers were evaluated on the basis of classification accuracy, or the total proportion of narratives which were correctly classified. The evaluation was performed using leave-one-out cross validation. In this procedure, one data point is left out, and the classifier is trained on the remaining data. The left-out data point can then be used as an unbiased test point. This procedure is repeated until each data point has been left out once, and the performance is averaged.

3. Results

We consider three separate classification tasks: (1) distinguishing between SD and controls; (2) distinguishing between PNFA and controls; and (3) distinguishing between SD and PNFA. The means and standard deviations for each attribute are compared in Tables 4–6. Group differences were measured using Welch's two-tailed, unpaired *t*-test, which does not assume that the two samples share the same variance. A significance level of $p < .05$ is indicated by a single asterisk, and $p < .01$ is indicated by a double asterisk. Because we were using the *t*-tests primarily as a means of feature selection, we did not adjust their significance levels for multiple comparisons. For each classification task, the set of significant features for that particular comparison formed the input vectors to the classifiers. The features were rescaled to have zero mean and unit variance before classification, to prevent features with large magnitudes (e.g., imageability) from dominating features with smaller magnitudes (e.g., fillers).

The features that were considered significant between the SD and control transcripts, and therefore used in that classification task, were: number of clauses, mean length of sentence, mean length of clause, T-units per sentence, total Yngve depth, mean Yngve depth, nouns, noun–verb ratio, noun ratio, demonstratives, adverbs, pronoun ratio, frequency, noun frequency, verb frequency, verb imageability, familiarity, noun familiarity, mean word length, and speech rate (see Table 4). For the task of classifying PNFA versus controls, the significant features were: number of words, T-units per sentence, demonstratives, frequency, verb frequency, age of acquisition, noun age of acquisition, mean word length, and speech rate (see Table 5). In the case of SD versus PNFA, only five features were significant: dependent clauses per clause, noun frequency, familiarity, noun familiarity, and occurrence of "um" (see Table 6).

The classification accuracies are given in Table 7. The baseline accuracies represent the accuracies that would be achieved by simply assigning every transcript to the larger of the two classes. That is, the baseline accuracy for SD ($n = 10$) versus controls ($n = 16$) would be achieved by simply classifying all the transcripts as controls ($16/26 = .615$). The two experimental scenarios involving patient groups versus control groups result in very high classification accuracies. The accuracies for classifying SD versus PNFA transcripts are not as high; however, they are well above the baseline for all three of the classifiers.

For comparison, we also evaluated the performance of the classifiers trained on all features as input, rather than just the features pre-selected by the *t*-tests. This method was expected to perform rather poorly due to overfitting. Although using more features may improve classification on the training set, it results in poor generalization to new data, as assessed with the cross-validation procedure. Indeed, performance in this case was lower than the results shown in Table 7: for SD versus controls, the accuracies ranged from .846 to .923, for PNFA versus controls the accuracies ranged from .700 to .800, and for SD versus PNFA they ranged from .625 to .667. This illustrates the necessity of feature selection prior to classifier training.

Because the classification takes place in high-dimensional feature space, it is difficult to visualize the models produced by the classifiers. Instead, it is useful to visualize the classes in two dimensions by using some form of dimensionality reduction. Here we use the method of partial least squares, or PLS (Haenlein and Kaplan, 2004). PLS is similar to the well-known method of principal components analysis, except that principal components analysis discovers the latent variables that best explain the variance in the attributes, while PLS discovers the latent variables that are most predictive of the response (in this case, the patient groups or class labels). In addition, PLS is appropriate when the number of attributes is high compared to the number of data points, which is the case here.

Scatter plots of the first two PLS components are shown in Fig. 1. Each of the plots show relatively good separation between the groups. We were interested to see whether two narratives that were located close together in the PLS plot shared some similarities, even when the participants who produced them were from different diagnostic groups (for example, the points labelled 1 and 2 in Fig. 1c). These transcripts are included in the Appendix, along with the transcripts associated with points 3 and 4 in Fig. 1c for the sake of comparison. Participants 1 and 2 have similar rates of speech, and Participant 1, although diagnosed with SD, makes several syntactic errors. Participant 2 was

Please cite this article in press as: Fraser KC, et al., Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2012.12.006

Table 4 – A comparison of SD and control features. Values shown are mean (standard deviation). Asterisks denote significance (*$p < .05$; **$p < .01$).

	Feature	SD	Controls	p Value	
1	Words	380.300 (272.429)	403.688 (121.380)	.8025	
2	Sentences	20.400 (13.550)	15.688 (6.877)	.3276	
3	T-units	.069 (.022)	.058 (.019)	.2036	
4	Clauses	.145 (.013)	.133 (.011)	.0292	*
5	Coordinate phrases	.028 (.010)	.029 (.007)	.6676	
6	Complex nominals	.098 (.039)	.083 (.020)	.2735	
7	Complex T-units	.037 (.009)	.030 (.009)	.0505	
8	Verb phrases	.177 (.020)	.167 (.015)	.2056	
9	Dependent clauses	.067 (.016)	.052 (.021)	.0543	
10	Mean length of sentence	20.135 (8.190)	28.608 (10.973)	.0346	*
11	Mean length of clause	7.035 (.603)	7.602 (.632)	.0329	*
12	Mean length of T-unit	16.217 (6.481)	19.465 (8.042)	.2701	
13	Dependent clauses per clause	.460 (.090)	.391 (.144)	.1459	
14	Dependent clauses per T-unit	1.094 (.497)	1.081 (.768)	.9571	
15	Verb phrases per T-unit	2.848 (1.276)	3.186 (1.196)	.5094	
16	Clauses per sentence	2.845 (1.010)	3.738 (1.242)	.0573	
17	Clauses per T-unit	2.292 (.786)	2.558 (1.007)	.4610	
18	Complex T-units per T-unit	.580 (.183)	.535 (.159)	.5353	
19	Coordinate phrases per T-unit	.463 (.334)	.589 (.369)	.3820	
20	Complex nominals per T-unit	1.558 (.720)	1.694 (1.082)	.7038	
21	T-units per sentence	1.250 (.230)	1.526 (.356)	.0249	*
22	Coordinate phrases per clause	.193 (.074)	.221 (.058)	.3150	
23	Complex nominals per clause	.673 (.229)	.627 (.162)	.5911	
24	Tree height	13.167 (2.359)	14.821 (2.401)	.0998	
25	Total Yngve depth	70.477 (41.393)	117.609 (72.438)	.0456	*
26	Maximum Yngve depth	5.091 (1.168)	6.023 (1.156)	.0613	
27	Mean Yngve depth	2.913 (.409)	3.345 (.498)	.0250	*
28	Nouns	.141 (.031)	.179 (.026)	.0051	**
29	Verbs	.207 (.028)	.200 (.019)	.4427	
30	Noun–verb ratio	.699 (.209)	.907 (.163)	.0164	*
31	Noun ratio	.403 (.072)	.472 (.047)	.0178	*
32	Inflected verbs	.635 (.127)	.706 (.086)	.1417	
33	Light verbs	.474 (.093)	.476 (.085)	.9527	
34	Determiners	.107 (.030)	.120 (.016)	.2203	
35	Demonstratives	.037 (.011)	.012 (.009)	.0000	**
36	Prepositions	.103 (.035)	.087 (.015)	.1994	
37	Adjectives	.034 (.013)	.038 (.009)	.3434	
38	Adverbs	.083 (.017)	.058 (.014)	.0010	**
39	Pronoun ratio	.508 (.094)	.416 (.068)	.0175	*
40	Function words	.467 (.033)	.453 (.033)	.2823	
41	Frequency	5.021 (.105)	4.803 (.104)	.0001	**
42	Noun frequency	3.861 (.231)	3.282 (.183)	.0000	**
43	Verb frequency	4.614 (.282)	4.378 (.184)	.0341	*
44	Imageability	477.721 (44.025)	507.025 (20.643)	.0729	
45	Noun imageability	560.959 (43.450)	580.710 (12.370)	.1913	
46	Verb imageability	416.117 (37.506)	385.947 (22.543)	.0387	*
47	Age of acquisition	258.881 (21.629)	257.814 (12.476)	.8894	
48	Noun age of acquisition	254.246 (33.465)	251.696 (19.006)	.8295	
49	Verb age of acquisition	260.465 (27.603)	266.521 (14.566)	.5338	
50	Familiarity	607.358 (12.903)	565.956 (10.052)	.0000	**
51	Noun familiarity	604.910 (20.954)	545.967 (17.119)	.0000	**
52	Verb familiarity	605.526 (19.573)	600.218 (13.388)	.4629	
53	Type-token ratio	.405 (.118)	.415 (.057)	.8028	
54	Mean word length	3.735 (.186)	3.997 (.152)	.0017	**
55	Fillers	.053 (.067)	.054 (.056)	.9876	
56	Um	.007 (.008)	.014 (.015)	.1613	
57	Uh	.046 (.061)	.040 (.060)	.8052	
58	Speech rate	104.048 (35.149)	160.779 (35.131)	.0007	**

diagnosed with PNFA, but tends to use high frequency words (such as *girls* instead of *stepsisters*). In contrast, the two transcripts from opposite sides of the PLS plots seem to be more clearly representative of their diagnostic groups.

Rather than analyze the factor loadings, which are not easily interpretable in PLS, we calculate *selectivity ratios*, which are closely related to the correlation between each attribute and the response (Kvalheim, 2010). A high

Please cite this article in press as: Fraser KC, et al., Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2012.12.006

Table 5 – A comparison of PNFA and control features. Values shown are mean (standard deviation). Asterisks denote significance (*$p < .05$; **$p < .01$).

	Feature	PNFA	Controls	p Value	
1	Words	302.214 (141.837)	403.688 (121.380)	.0466	*
2	Sentences	16.143 (12.526)	15.688 (6.877)	.9049	
3	T-units	.061 (.023)	.058 (.019)	.7698	
4	Clauses	.141 (.020)	.133 (.011)	.2027	
5	Coordinate phrases	.028 (.013)	.029 (.007)	.7585	
6	Complex nominals	.092 (.017)	.083 (.020)	.2005	
7	Complex T-units	.030 (.010)	.030 (.009)	.9345	
8	Verb phrases	.167 (.017)	.167 (.015)	.9699	
9	Dependent clauses	.055 (.016)	.052 (.021)	.7312	
10	Mean length of sentence	24.501 (12.192)	28.608 (10.973)	.3437	
11	Mean length of clause	7.299 (.986)	7.602 (.632)	.3348	
12	Mean length of T-unit	19.655 (8.814)	19.465 (8.042)	.9516	
13	Dependent clauses per clause	.384 (.082)	.391 (.144)	.1333	
14	Dependent clauses per T-unit	1.054 (.545)	1.081 (.768)	.9106	
15	Verb phrases per T-unit	3.212 (1.400)	3.186 (1.196)	.9577	
16	Clauses per sentence	3.346 (1.493)	3.738 (1.242)	.4444	
17	Clauses per T-unit	2.716 (1.243)	2.558 (1.007)	.7078	
18	Complex T-units per T-unit	.517 (.146)	.535 (.159)	.7456	
19	Coordinate phrases per T-unit	.615 (.521)	.589 (.369)	.8763	
20	Complex nominals per T-unit	1.767 (.790)	1.694 (1.082)	.8331	
21	T-units per sentence	1.256 (.254)	1.526 (.356)	.0232	*
22	Coordinate phrases per clause	.203 (.093)	.221 (.058)	.5336	
23	Complex nominals per clause	.660 (.130)	.627 (.162)	.5493	
24	Tree height	12.933 (2.718)	14.821 (2.401)	.0555	
25	Total Yngve depth	108.405 (78.010)	117.609 (72.438)	.7415	
26	Maximum Yngve depth	5.275 (1.364)	6.023 (1.156)	.1197	
27	Mean Yngve depth	3.085 (.589)	3.345 (.498)	.2070	
28	Nouns	.155 (.040)	.179 (.026)	.0644	
29	Verbs	.191 (.025)	.200 (.019)	.2804	
30	Noun–verb ratio	.837 (.286)	.907 (.163)	.4276	
31	Noun ratio	.444 (.082)	.472 (.047)	.2775	
32	Inflected verbs	.706 (.096)	.706 (.086)	.9923	
33	Light verbs	.538 (.141)	.476 (.085)	.1666	
34	Determiners	.130 (.032)	.120 (.016)	.3183	
35	Demonstratives	.026 (.018)	.012 (.009)	.0210	*
36	Prepositions	.088 (.032)	.087 (.015)	.8884	
37	Adjectives	.030 (.017)	.038 (.009)	.1219	
38	Adverbs	.069 (.030)	.058 (.014)	.2253	
39	Pronoun ratio	.476 (.095)	.416 (.068)	.0601	
40	Function words	.478 (.045)	.453 (.033)	.0968	
41	Frequency	4.962 (.118)	4.803 (.104)	.0006	**
42	Noun frequency	3.451 (.317)	3.282 (.183)	.0936	
43	Verb frequency	4.608 (.237)	4.378 (.184)	.0072	**
44	Imageability	509.119 (40.552)	507.025 (20.643)	.8634	
45	Noun imageability	579.078 (23.669)	580.710 (12.370)	.8192	
46	Verb imageability	404.073 (49.196)	385.947 (22.543)	.2215	
47	Age of acquisition	245.879 (14.862)	257.814 (12.476)	.0260	*
48	Noun age of acquisition	235.038 (20.937)	251.696 (19.006)	.0316	*
49	Verb age of acquisition	264.400 (13.607)	266.521 (14.566)	.6834	
50	Familiarity	573.625 (18.408)	565.956 (10.052)	.1808	
51	Noun familiarity	561.110 (24.689)	545.967 (17.119)	.0668	
52	Verb familiarity	589.838 (19.242)	600.218 (13.388)	.1043	
53	Type-token ratio	.421 (.046)	.415 (.057)	.7421	
54	Mean word length	3.769 (.136)	3.997 (.152)	.0002	**
55	Fillers	.083 (.080)	.054 (.056)	.2584	
56	Um	.025 (.027)	.014 (.015)	.1948	
57	Uh	.058 (.084)	.040 (.060)	.5937	
58	Speech rate	78.468 (27.978)	160.779 (35.131)	.0000	**

selectivity ratio indicates that a feature is very influential with respect to the response. The details of calculating this measure are given by Kvalheim (2010); we used a pre-existing Matlab package to perform the analysis (Li, 2011). Fig. 2 shows the selectivity ratios for each feature in the three cases. For ease of interpretation, each feature with a selectivity ratio of greater than a cut-off of .5 is given for each of the three classification problems. In the majority of

Please cite this article in press as: Fraser KC, et al., Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2012.12.006

Table 6 – A comparison of SD and PNFA features. Values shown are mean (standard deviation). Asterisks denote significance (*$p < .05$; **$p < .01$).

	Feature	SD.	PNFA	p Value	
1	Words	380.300 (272.429)	302.214 (141.837)	.4223	
2	Sentences	20.400 (13.550)	16.143 (12.526)	.4436	
3	T-units	.069 (.022)	.061 (.023)	.3518	
4	Clauses	.145 (.013)	.141 (.020)	.6028	
5	Coordinate phrases	.028 (.010)	.028 (.013)	.9334	
6	Complex nominals	.098 (.039)	.092 (.017)	.6376	
7	Complex T-units	.037 (.009)	.030 (.010)	.0580	
8	Verb phrases	.177 (.020)	.167 (.017)	.2393	
9	Dependent clauses	.067 (.016)	.055 (.016)	.0805	
10	Mean length of sentence	20.135 (8.190)	24.501 (12.192)	.3056	
11	Mean length of clause	7.035 (.603)	7.299 (.986)	.4254	
12	Mean length of T-unit	16.217 (6.481)	19.655 (8.814)	.2829	
13	Dependent clauses per clause	.460 (.090)	.384 (.082)	.0472	*
14	Dependent clauses per T-unit	1.094 (.497)	1.054 (.545)	.8509	
15	Verb phrases per T-unit	2.848 (1.276)	3.212 (1.400)	.5161	
16	Clauses per sentence	2.845 (1.010)	3.346 (1.493)	.3381	
17	Clauses per T-unit	2.292 (.786)	2.716 (1.243)	.3188	
18	Complex T-units per T-unit	.580 (.183)	.517 (.146)	.3825	
19	Coordinate phrases per T-unit	.463 (.334)	.615 (.521)	.3953	
20	Complex nominals per T-unit	1.558 (.720)	1.767 (.790)	.5084	
21	T-units per sentence	1.250 (.230)	1.256 (.254)	.9541	
22	Coordinate phrases per clause	.193 (.074)	.203 (.093)	.7648	
23	Complex nominals per clause	.673 (.229)	.660 (.130)	.8714	
24	Tree height	13.167 (2.359)	12.933 (2.718)	.8246	
25	Total Yngve depth	70.477 (41.393)	108.405 (78.010)	.1386	
26	Maximum Yngve depth	5.091 (1.168)	5.275 (1.364)	.7270	
27	Mean Yngve depth	2.913 (.409)	3.085 (.589)	.4071	
28	Nouns	.141 (.031)	.155 (.040)	.3591	
29	Verbs	.207 (.028)	.191 (.025)	.1440	
30	Noun–verb ratio	.699 (.209)	.837 (.286)	.1844	
31	Noun ratio	.403 (.072)	.444 (.082)	.2122	
32	Inflected verbs	.635 (.127)	.706 (.096)	.1550	
33	Light verbs	.474 (.093)	.538 (.141)	.1935	
34	Determiners	.107 (.030)	.130 (.032)	.0862	
35	Demonstratives	.037 (.011)	.026 (.018)	.0688	
36	Prepositions	.103 (.035)	.088 (.032)	.3086	
37	Adjectives	.034 (.013)	.030 (.017)	.5577	
38	Adverbs	.083 (.017)	.069 (.030)	.1586	
39	Pronoun ratio	.508 (.094)	.476 (.095)	.4313	
40	Function words	.467 (.033)	.478 (.045)	.5254	
41	Frequency	5.021 (.105)	4.962 (.118)	.2139	
42	Noun frequency	3.861 (.231)	3.451 (.317)	.0014	**
43	Verb frequency	4.614 (.282)	4.608 (.237)	.9557	
44	Imageability	477.721 (44.025)	509.119 (40.552)	.0916	
45	Noun imageability	560.959 (43.450)	579.078 (23.669)	.2527	
46	Verb imageability	416.117 (37.506)	404.073 (49.196)	.5036	
47	Age of acquisition	258.881 (21.629)	245.879 (14.862)	.1211	
48	Noun age of acquisition	254.246 (33.465)	235.038 (20.937)	.1309	
49	Verb age of acquisition	260.465 (27.603)	264.400 (13.607)	.6845	
50	Familiarity	607.358 (12.903)	573.625 (18.408)	.0000	**
51	Noun familiarity	604.910 (20.954)	561.110 (24.689)	.0001	**
52	Verb familiarity	605.526 (19.573)	589.838 (19.242)	.0659	
53	Type-token ratio	.405 (.118)	.421 (.046)	.6828	
54	Mean word length	3.735 (.186)	3.769 (.136)	.6341	
55	Fillers	.053 (.067)	.083 (.080)	.3290	
56	Um	.007 (.008)	.025 (.027)	.0335	*
57	Uh	.046 (.061)	.058 (.084)	.6864	
58	Speech rate	104.048 (35.149)	78.468 (27.978)	.0736	

cases, these are the same features which were found to be significant and included in the classifiers above. We expect there to be some discrepancies, as the individual t-tests do not take into account correlations between variables, while the PLS analysis does. Selectivity ratios for influential features may be reduced in the presence of a second feature highly correlated with the first, as the two share variance that predicts group membership.

Please cite this article in press as: Fraser KC, et al., Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2012.12.006

Table 7 – Accuracies for the three classifiers, compared to a simple baseline classifier.

	SD versus control	PNFA versus control	SD versus PNFA
Baseline	.615	.533	.583
Naïve Bayes	.923	.900	.792
Logistic regression	.962	.933	.708
SVM	1.00	.967	.750

4. Discussion

In this study we set out to determine whether computational methods could reliably distinguish between healthy controls and patients with PPA, as well as differentiate the two patient groups, based upon samples of narrative speech. We found that even with relatively short samples of narrative speech (i.e., for machine learning purposes), classifiers were able to achieve this goal with a high degree of accuracy. In addition, we wished to determine how the automatically extracted features compared to previous findings in the literature with respect to these two subtypes. In general, we found that our procedure identified many of the same features that have been previously noted to differ between groups (e.g., word frequency, speech rate, demonstrative pronouns), with some surprising findings (e.g., the lack of syntactic features as differentiating ones) and some new features identified (e.g., adverbs and word length). We discuss these issues below.

4.1. *Classification*

Our results show that machine learning classifiers can distinguish between controls and each of the two patient groups, SD and PNFA, with a high degree of accuracy. Although less accurate than in comparison to controls, they also distinguish well between the two patient groups. The performance of the classifiers varied across the three tasks: SVM achieved the highest accuracy for SD versus controls and PNFA versus controls, while naïve Bayes performed best for SD versus PNFA. Logistic regression achieved the second-highest accuracy in the first two cases, but was the worst at distinguishing between SD and PNFA. We also note the relatively high accuracy of naïve Bayes despite obvious correlations between the features in some cases.

4.2. *Features that distinguished the groups and comparison with previous findings on PPA*

The PLS analysis identified the features that best predicted group membership. The selectivity ratios, together with the group means on each feature, provided valuable information on the characteristics of narrative speech in each group.

The features that best distinguished the SD patients from controls were higher familiarity and frequency of words (particularly nouns), increased production of adverbs and demonstratives, production of shorter clauses, and reduced word length and speech rate. The familiarity of the words produced, which was greater in SD, was the feature with the highest selectivity ratio. The familiarity of nouns in particular

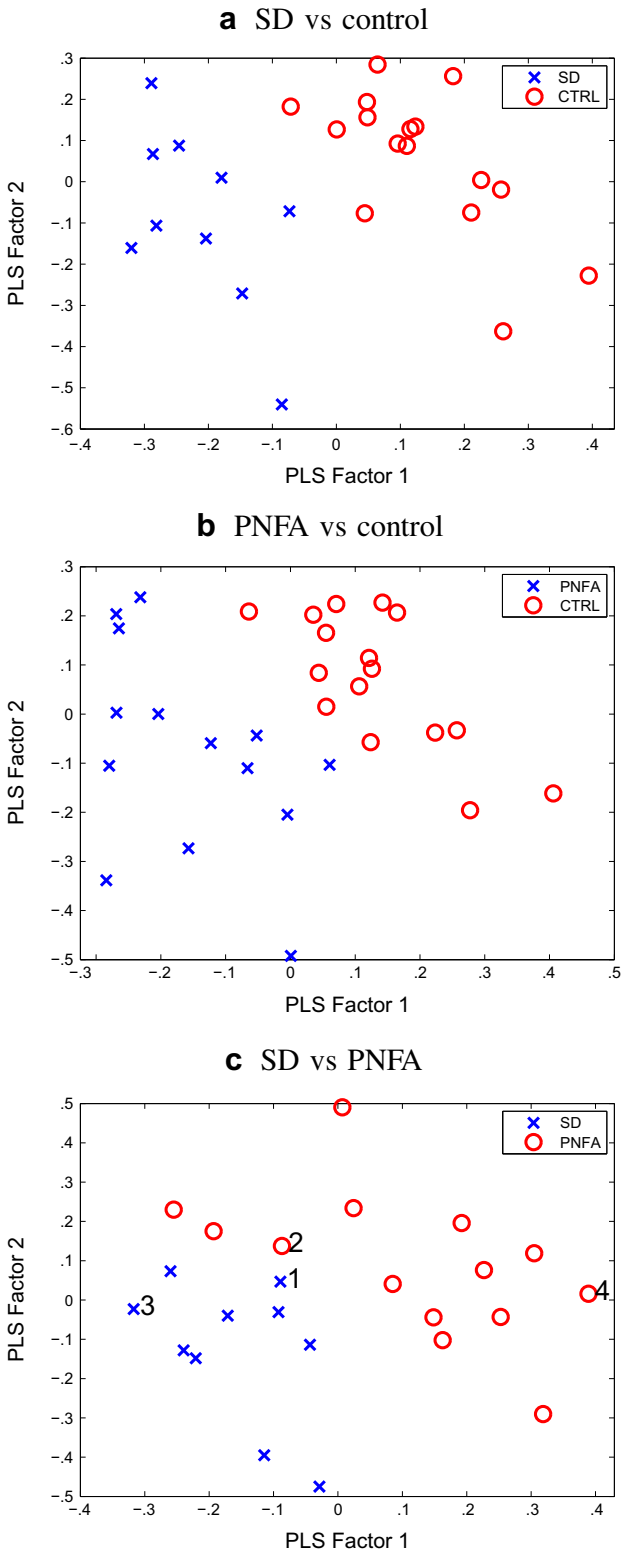
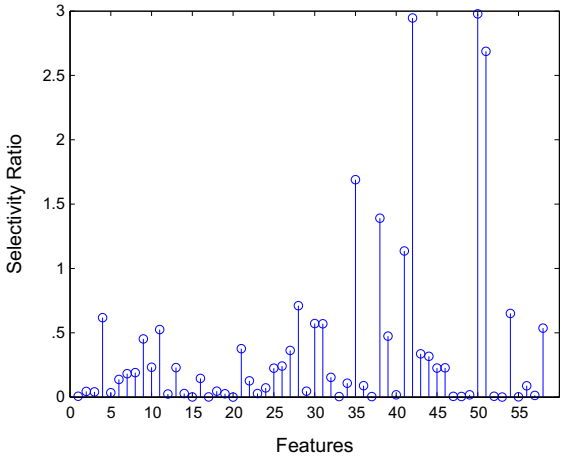


Fig. 1 – PLS analysis of the data, with each point representing one transcript. Transcriptions for the participants labelled in (c) are provided in the Appendix.

was also ranked highly. The finding that the SD patients produced words with higher familiarity ratings than controls is consistent with findings from studies of naming (Lambon Ralph et al., 1998; Woollams et al., 2008), but to the best of

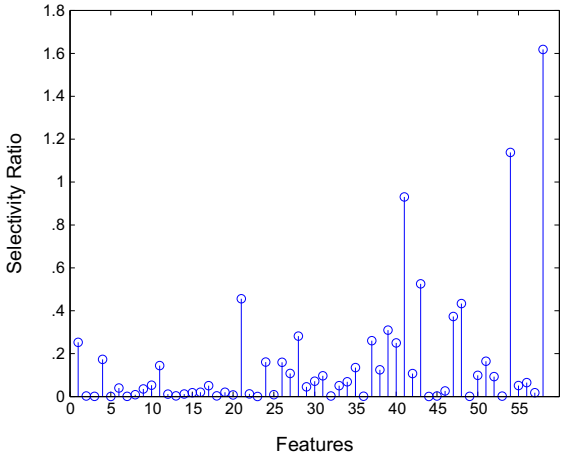
Please cite this article in press as: Fraser KC, et al., Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2012.12.006

a SD vs. control



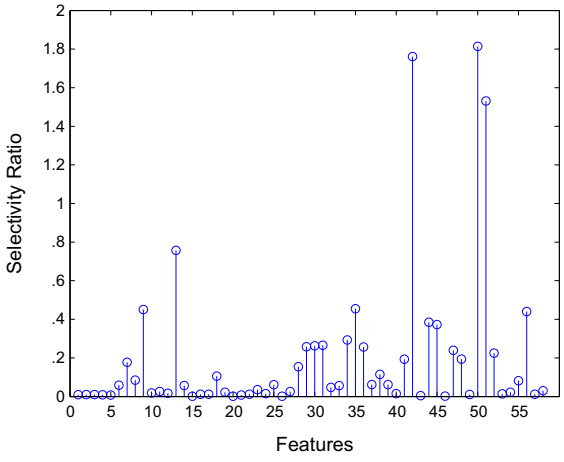
Feature		SR
50	familiarity	2.98
42	noun frequency	2.95
51	noun familiarity	2.69
35	demonstratives	1.69
38	adverbs	1.39
41	frequency	1.13
28	nouns	.71
54	word length	.65
4	clauses	.62
30	noun-verb ratio	.57
31	noun ratio	.57
58	speech rate	.54
11	mean clause length	.53

b PNFA vs. control



Feature		SR
58	speech rate	1.62
54	word length	1.14
41	frequency	.93
43	verb frequency	.53

c SD vs. PNFA



Feature		SR
50	familiarity	1.81
42	noun frequency	1.76
51	noun familiarity	1.53
13	dependent clauses per clause	.76

Fig. 2 – PLS selectivity ratio for each of the features. On the left, plots of the selectivity ratio for every feature. The feature numbers on the horizontal axes refer to the feature numbers in Table 2. Note the difference in scale on the vertical axes. On the right, the features with selectivity ratio greater than .5.

our knowledge has not been previously documented in connected speech. It also fits with previous research demonstrating that SD patients' semantic knowledge is best preserved for familiar items (Funnel, 1996; Simons et al., 2001).

The SD patients tended to use higher frequency nouns than the controls, and this feature had the second highest selectivity ratio. Overall word frequency also distinguished well between the groups and these findings demonstrate the

Please cite this article in press as: Fraser KC, et al., Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2012.12.006

robust effect of frequency on language production in SD. It has been established that frequency has a pervasive influence on naming in SD (Lambon Ralph et al., 1998; Woollams et al., 2008), but this influence is less well documented in connected speech. Two studies which used picture description tasks observed that SD patients produced higher frequency nouns than controls (Bird et al., 2000; Wilson et al., 2010). Meteyard and Patterson (2009) did not evaluate frequency per se, but their analysis of structured interviews showed that patients with SD tended to replace content words with higher frequency (and less specific) words. We have found that patients with SD use higher frequency words overall, and that nouns are particularly affected.

Familiarity and frequency, the features with the highest selectivity ratios, are correlated with each other (Tanaka-Ishii and Terada, 2011). It is interesting to note that individually they would have had even higher selectivity ratios if only one of the features had been included in the model.

The SD patients produced more demonstratives and adverbs than controls, and this distinguished the groups. The increased reliance on demonstrative pronouns, which comprise the words *these*, *those*, *this*, *that*, *here*, and *there*, may reflect the tendency of SD patients to make substitutions of less specific words (Meteyard and Patterson, 2009) and use vague terms (Kavé et al., 2007; Patterson and MacDonald, 2006); the automated analysis techniques used here did not enable us to evaluate whether these terms had clear referents when used. Previous studies have documented over-reliance on pronouns in SD (Kavé et al., 2007; Meteyard and Patterson, 2009; Patterson and MacDonald, 2006; Wilson et al., 2010), but to the best of our knowledge this is the first examination specifically of demonstrative pronouns. The increased use of adverbs may at least partially reflect that fact that some participants in this patient group started many of the sentences in their narratives with *then* or *so*, sometimes preceded by *and*. Because *then* and *so* are classified as adverbs (describing when or why something happened), this inflates the count on these words. Repeated use of the same syntactic structure is compatible with the idea that SD patients are unable to produce the full range of syntactic structures (Benedet et al., 2006), but needs to be investigated further before firm conclusions can be drawn.

The SD patients produced fewer nouns than the controls, and reduced noun production accounted for 3 of the features which had high selectivity ratios (nouns, noun-to-verb ratio, and noun ratio). Difficulty with production of nouns is an expected finding. Impaired confrontation naming is a core diagnostic feature in SD (Gorno-Tempini et al., 2011), and is expected to lead to corresponding problems with word-finding in connected speech (see Sajjadi et al., 2012). Moreover, previous studies have documented reduced production of nouns in the connected speech of people with SD (Bird et al., 2000; Ash et al., 2009; Patterson and MacDonald, 2006; Kavé et al., 2007). Bird et al. (2000) contrasted noun and verb production by SD patients on a picture description task and provided evidence that production of nouns was more affected as a result of their lower relative frequency. This is compatible with the current results in that our SD patients produced higher frequency, and proportionally fewer, nouns than controls (and these features had high selectivity ratios).

The mean word length for the SD patients was slightly shorter than for controls, and this feature distinguished well between the groups. The effect of word length has not been examined in connected speech. We attribute this small but significant effect of length to availability of words, rather than to difficulty with pronouncing long words. Many of the long words used by controls were used less often by the SD patients. The most frequently used long word for both groups was *Cinderella*, which was used a total of 69 times by the controls and a total of 17 times by the SD patients (an average of 4.3 *vs* 1.7 times per narrative). Other long words that were used more often by controls than patients include, for example, *slipper* which was used 48 times by controls but never by SD patients, and *beautiful* which was used 25 times by controls and 5 times by SD patients (an average of 1.6 versus .5 times per narrative).

The number, and mean length, of clauses both distinguished between the SD and control groups, although the numerical differences were rather small. The SD patients produced more clauses than the controls, but on average their clauses were shorter. The explanation for this is not clear. It seems unlikely to be an artefact of the automatic coding, as the clause counts had high agreement with the human annotator, even with the sentence-by-sentence comparison. It may be associated with the reduced production of nouns, which could result in fewer nouns per clause. In addition, inspection of the transcriptions indicates that the SD patients were more likely to produce "filler" comments such as *you know* and *whatever you call it*, which are relatively short clauses. We know of no identical analyses in other studies, although some have used similar measures: Patterson and MacDonald (2006) found that SD and controls produced similar numbers of clauses, while Sajjadi et al. (2012) found that SD patients produced reduced proportions of syntactically complex clauses relative to controls (data on simple clauses were not reported). Further work would be required to understand the basis and potential significance of the present finding that SD patients tend to produce shorter, but more, clauses than controls.

The final feature which distinguished well between the SD patients and controls was speech rate. As noted in the Introduction, findings with respect to speech rate in SD have been inconsistent. The slower rate for SD patients may be due to pauses in speech while searching for a word, and seems unlikely to reflect a motor speech problem. Wilson et al. (2010) also documented slower speech rate in SD patients than in controls, but found that the maximum speech rate (which they defined as the three most rapidly spoken sequences of ten or more words) for their patients was normal; they suggested that the slower rate reflected impairment in higher-level processes, and we concur with this idea.

The features that distinguish SD patients from controls can largely be attributed to the semantic memory impairment. This seems to be the dominant influence in the language output of this group, and can account for the increased reliance on more familiar and frequent words, use of general terms such as demonstratives and adverbs like *then* and *so*, reduced production of nouns, and pauses for word-finding (which could explain the relatively slower speech rate).

Please cite this article in press as: Fraser KC, et al., Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2012.12.006

The features that distinguished the PNFA patients from controls were reduced speech rate and word length, as well as higher frequency of words and of verbs in particular. Not surprisingly, slower speech rate was the feature that best distinguished PNFA from controls. A reduced rate of speech is one of the diagnostic features (Gorno-Tempini et al., 2004), and has been documented in other studies of connected speech production in PNFA (Ash et al., 2010, 2006; Graham et al., 2004; Knibb et al., 2009; Thompson et al., 2012).

The PNFA patients tended to produce shorter words than controls, and this feature had a high selectivity ratio. Increased word length of stimulus items has been shown to deleteriously affect naming, reading and repetition in PNFA (Croot et al., 1998; Graham et al., 2004), and this effect could arise from phonological or motor speech impairment(s). In the case of narrative speech, it could also arise from word-finding difficulties, which would affect availability of words (as suggested for the SD patients). The current analyses do not inform the choice of explanation, and indeed the use of shorter words could arise from different causes in different individuals.

Like the SD patients, the PNFA patients were distinguished from controls by overall word frequency. For the PNFA patients (versus controls), verb frequency also had a high selectivity ratio. Once again we have a situation where two correlated features both have high selectivity ratios, suggesting that each would have had an even higher ratio if they had been included in the model without the other. Studies have shown that naming of verbs/actions in PNFA is more impaired than naming of nouns/objects (Cotelli et al., 2006; Hillis et al., 2004). Although we do not know of a study which assesses the effect of frequency on naming of verbs in PNFA, our results suggest that production of higher frequency verbs is better preserved. The finding that the PNFA patients produce higher frequency verbs than the controls cannot be due to excessive use of light verbs (which tend to be high in frequency), because the *t*-tests indicate no difference between PNFA and controls in use of light verbs. Despite this apparent difficulty with verb production, the PNFA patients in this study produced nouns and verbs in normal proportions as demonstrated by equivalent noun-to-verb ratios for PNFA patients and controls. An increase in noun-to-verb ratio is taken to indicate difficulty with verb production, and has been reported in other studies of connected speech in PNFA (Thompson et al., 1997). Consistent with the current findings, Graham et al. (2004) also documented normal noun-to-verb ratios but suggested that the verbs produced by PNFA patients were less specific than those produced by controls.

The results also yielded interesting findings regarding the features that distinguished between the two patient groups. SD patients used nouns which were more frequent and familiar, and words which were more familiar overall, than PNFA patients. These features are identical to those that best discriminated between SD patients and controls, and have the same rank order with respect to their selectivity ratios. As noted above, frequency and familiarity are known to affect naming performance in SD (Woollams et al., 2008; Lambon Ralph et al., 1998), and clearly these factors affect word-finding in connected speech as well. The greater impact (relative to PNFA) of familiarity and frequency upon the speech of the SD patients may be due to their greater semantic impairment, or to the fact that they are more anomic than the PNFA patients (as documented in Table 1).

The final feature that differentiated the two patient groups was the number of dependent clauses per clause. This indicates that the SD patients produce a proportionally greater number of clauses which could not form sentences on their own. In their study of connected speech in SD, Meteyard and Patterson (2009) documented increased production of restarts, which they defined as a sentence which was incomplete and then started again. We did not count restarts, but inspection of the transcriptions reveals that there were many, and this could account at least partially for the increased production of incomplete (i.e., dependent) clauses. The finding that production of dependent clauses was greater in SD than PPA should, however, be regarded as tentative as there is uncertainty (noted above) in the ability of the system to properly identify dependent clauses. Further work would be needed to clarify the reliability and interpretation of this result.

Some surprising findings also emerged. While it is well documented that PNFA patients tend to have sparse output, producing fewer words than either controls or SD patients (Graham et al., 2004; Wilson et al., 2010), total word count did not emerge as an important feature distinguishing the groups in the PLS plots. It is also surprising that so few of the features measuring syntactic complexity emerged as main distinguishing features, particularly for the PNFA group. This may have occurred because, as noted in the Methods section, not all of the PNFA patients exhibited agrammatism. Alternatively, it may be due to the way the syntactic analyses were performed. As mentioned in Section 2.3, Lu's syntactic complexity analyzer was originally designed to be applied to written documents by second-language learners, and so may not be ideally suited to the analysis of aphasic speech. Our goal was to test its performance in this domain, and although we found that it was effective in detecting clause boundaries, and returned counts that were highly correlated with manual counts, it did encounter particular difficulty in labelling clauses as dependent or independent. In part, this may be attributable to the uncertainty about sentence boundaries as determined by a human transcriber, as opposed to quantitative linguistic criteria. We note that sentence boundaries are frequently ambiguous in natural speech, aphasic or otherwise. Further methodological development of automated analysis for syntactic complexity is a promising avenue for future research, particularly if it can be made to operate mainly within rather than across clause boundaries.

To summarize, the automated analyses indicated that SD patients showed an over-reliance on words which were high in familiarity and/or frequency, and this applied particularly to nouns. They also produced proportionally fewer nouns, but more demonstratives (e.g., *this*, *these*) and more adverbs (e.g., *so*, *then*). In contrast, the speech of the PNFA patients was characterized by reduced speech rate and word length; this group also tended to use words which were high in frequency and this applied particularly to verbs. Verbs were, however, produced in normal proportions. The SD patients were distinguished from PNFA by their relatively greater use of words with higher familiarity and frequency.

Please cite this article in press as: Fraser KC, et al., Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2012.12.006

4.3. *Future directions*

In this work, we have demonstrated that fairly high classification accuracies can be achieved through automated quantitative analysis of speech samples, with relatively little human intervention required except for transcription. Given that procedures already exist for diagnosing PPA (Gorno-Tempini et al., 2011), one might ask what added value there is in developing an automated classification approach based on naturalistic speech alone. In fact, a diagnostic classifier provides a starting point for a much more extensive use of speech samples in applications beyond diagnosis. The ultimate goal of research into language disorders is to develop techniques to intervene effectively, either to restore function, or to slow its decline. Many diagnostic tests are not generally well suited for longitudinal assessment of language function, due to practice and familiarity effects. Furthermore, a patient's speech may change in significant ways that are not necessarily reflected by formal tests, but can be nonetheless captured and quantified by linguistic analysis. The automated measurement of many parameters provides the maximum opportunity to reveal significant changes across time, and the use of these parameters in classification provides a means to decide which parameters are most indicative of the function of underlying language systems.

We focused here on differential diagnosis between SD and PNFA, due to their association with degeneration in distinct brain regions. We consider SD and PNFA to be good models for studying dysfunction in the ventral and dorsal language networks, respectively. However, the same methods can be applied to tracking language dysfunction in a variety of neural disorders, including AD. Language symptoms are relatively common in AD, but highly variable across individuals (Taler and Phillips, 2008). This high variability is presumably due to differential spread of cortical pathology (Stopford et al., 2008), in contrast to the medial temporal lobe pathology that is more universal in AD and underlies the disease's characteristic episodic memory impairment. As greater cortical involvement accompanies the progression of AD (Singh et al., 2006), quantitative analyses of speech content may provide a sensitive measure of disease severity, useful for the evaluation of interventions designed to slow the progression. Similarly, this analysis may be useful in contexts where the goal is to bring about improvement rather than slow decline, such as in rehabilitation of post-stroke aphasia. Although narrative speech is widely regarded as a rich and ecologically valid source of information about linguistic function, the labour-intensive nature of its analysis has precluded its widespread adoption in research and clinical practice. Therefore, the analysis methods presented here, which can be fully automated beyond the transcription stage, may offer a basis for routine incorporation of narrative speech into cognitive evaluation for a wide range of disorders beyond PPA.

Longitudinal evaluation of language decline will require extensions to our present approach of binary classification. The binary classifiers presented here are aimed at categorizing an individual patient in an all-or-none fashion, and the metric of their success at present is the rate at which their classification matches the clinical diagnosis. However, some classifiers (e.g., naïve Bayes and logistic regression, out of the ones considered here) output a probabilistic estimate of class membership for each individual case. Such values could be tracked for an individual over time, based on a classifier trained on a static set of observations in different patients. Alternatively, instead of classification, one could use machine learning techniques based on regression, which seek to map continuous input variables to continuous output variables. Many such techniques are available, including ridge regression (Hoerl and Kennard, 1970), support vector regression, and relevance vector regression (Tipping, 2001). These techniques have become popular in neuroimaging, as investigators have sought to reveal relationships between continuous behavioural variables and multi-voxel measures of brain structure or activity (for review, see Cohen et al., 2011). In the case of PPA, these continuous coding techniques could be used to determine quantitative relationships between aspects of speech and patterns of brain atrophy or hypoactivity. As large datasets containing both speech data and neuroimaging measures from the same participants become available, we expect these machine learning methods to play an increasingly large role in elucidating the neural bases of language processing in both health and disease.

Acknowledgements

This research was supported by a grant from the Canadian Institutes of Health Research (CIHR), grant # MOP-82744 to E.R., S.E.B., C.L., and N.G., by a grant from the Natural Sciences and Engineering Research Council (NSERC) to G.H., and by a grant from the Alzheimer's Association to J.M. The authors also acknowledge the support of Toronto Rehabilitation Institute, which receives funding under the Provincial Rehabilitation Research Program from the Ministry of Health and Long-Term Care in Ontario. Additional support was provided by the Heart and Stroke Foundation Centre for Stroke Recovery. We thank Lily Panamsky for performing the manual parsing and tagging. We thank Dr. Karalyn Patterson and an anonymous reviewer for their helpful comments. We especially thank the participants for their patience and perseverance, and we thank Dr. David Tang-Wai and Dr. Tiffany Chow for referrals to the study.

Appendix A. Sample transcriptions

Transcriptions of four participants' narratives of the Cinderella story are provided below. The transcriptions are numbered to reflect their location on the plot in Fig. 1c. Transcriptions 1 and 2 are from the participants in each patient group who were closest to each other in the PLS analysis comparing SD versus NFPA; their proximity in the PLS plot suggests that they share some similarities despite belonging to different patient groups. Examples 3 and 4 were located on opposite sides of the plot, which suggests they may be more typical examples of SD and PNFA speech, respectively.

The following annotations were used (although all annotations are removed before computational analysis, as are commas):

(hhh)	laughter, exhalation pulse
:	elongated vowel/consonant
(X)	X represents length of pause in seconds
[###]	unintelligible
((text))	gloss/transcriber comment
/text/	phonological transcription

Transcription 1 (SD, patient 29). Speech rate: 116.3 words per minute.

Well she was a kind of, a uh not a woman there, she was a bad woman. She had a couple of kids, women, girls, and he she had her too. And she uh, she's uh had to do all the work, do the washing and uh everyday she's at work and they don't have to work at all. And she used to go around tell her to do everything for them. So anyway uh [###] sort on it because uh then they came about this uh dance, and uh she wouldn't let her go. Two girls would go, but not her. But anyway, she got a fairy. Found a fairy a woman talked to her and that. And she got sitting up and they said, she got him to there with all dressed up to, uh I don't know the mice now, they helped her a lot apparently, but I don't know that too much about it, but they did. And she got real nice, she got dressed up she looked for every the stuff and got to the feet, the feet yeah the the shoes. And she went there with it, and when he went in, she uh uh what happened she drove ahead of it, and ran out, go away, and she dropped the (1) shoe. So anyway, this fellow came looking out, found you know, one of these fair fellows the big fellows, he went for the prince. [###] And he had to find that woman, I want to find that, and then he get this, he got the leg, the foot, the shoe. So he went over to everybody and when he found her, just when she found out and and and the prince took her away.

Transcription 2 (PNFA, patient 41). Speech rate: 110.5 words per minute.

OK. So, Cinderella, one day she ended up in, the middle of, um, in the in the in the, inside. So she was, she went in: to this house. And, sh: what basically what she had to do was she had to go down on her hands and knees and she had to clean. That was her job, um and, she also had, two or three, um, three girls, where they didn't have to do anything, which was interesting. And they were all excited about going to the ball. Um:, in the meantime, she, she also has a, grandmother, who looks af-, who looked after her. And then finally, she did actually, well first of all, she realized that the three people, they were her sisters as such, and then there was this other lady who basically was not a nice person at all. And: so that went on for quite a while, so they were, it was all about them and not about her, and then she actually met, her, beautiful person.

Transcription 3 (SD, patient 2). Speech rate: 101.5 words per minute.

Well, Cinderella, um (2) came to um, the co- n- I gon' say cottage but it's not a cottage obviously it's a mm quite a p- a place. And um, she uh initially is not th:ought of too much and uh the little animals sort of thing are b- uh down there and uh and as a matter of fact uh they uh come to like her as a matter of fact too. And um she uh is uh going there and these little things are going along and uh the uh older lady and so on uh not th- that g- good to her ah I I suspect. And then there's a shellow ((fellow)) who uh looks quite nice and he thinks that she looks quite nice and it ends up to some degree when they eventually get married. That's all I can think of right there, but sorry. (2) Uh (2) and there's all all s- uh sort of little little things that are running around and uh and that but part of the thing but she doesn't br- b- bother them but she rather likes that the gentleman who is probably one of the king's dau- s- s- sons. And away they go. Anyways that's (hhh) just looking at it slightly like that. I'm sure I missed a lot of things but nevertheless that's positive and I'm glad she's fi- having fun. And so are the (hhh) little things too. Anyways, there.

Transcription 4 (PNFA, patient 31). Speech rate: 54.6 words per minute.

Okay. Um, Cinderella, they have uh a stepfathers and the uh godmother or whatever. Um (4) uh um (11) I, I guess they're going to go to the ball and then um all those mice they're going to do the dress (hhh) and um (3), and u:m oh the pumpkin (hhh) and uh (3) well the godmother will do the pumpkin and um (4), and then she'll go to the ball (3) and then (4) he has, he has a (3) him (hhh) him um anyways, so anyways twelve o'clock she'll go down the stairs. And she has a glass slipper, but it's, it's not it's just in the stairs, so anyways she um she had a pumpkin. She went (2) um went to her castle again and then um (3) and then they had a glass slipper for her and then: the prince and the Cinderella.

REFERENCES

Ash S, McMillan C, Gunawardena D, Avants B, Morgan B, Khan A, et al. Speech errors in progressive non-fluent aphasia. *Brain and Language*, 113(1): 13–20, 2010.

Ash S, Moore P, Antani S, McCawley G, Work M, and Grossman M. Trying to tell a tale: Discourse impairments in progressive aphasia and frontotemporal dementia. *Neurology*, 66(9): 1405–1413, 2006.

Ash S, Moore P, Vesely L, Gunawardena D, McMillan C, Anderson C, et al. Non-fluent speech in frontotemporal lobar degeneration. *Journal of Neurolinguistics*, 22(4): 370–383, 2009.

Benedet M, Patterson K, Gomez-Pastor I, and Luisa Garcia de la Rocha M. 'Non-semantic' aspects of language in semantic dementia: As normal as they're said to be? *Neurocase*, 12(1): 15–26, 2006.

Berndt RS, Wayland S, Rochon E, Saffran E, and Schwartz M. *Quantitative Production Analysis: A Training Manual for the Analysis of Aphasic Sentence Production*. Hove, UK: Psychology Press, 2000.

Bird H, Lambon Ralph MA, Patterson K, and Hodges JR. The rise and fall of frequency and imageability: Noun and verb production in semantic dementia. *Brain and Language*, 73: 17–49, 2000.

Bishop DVM. *Test for the Reception of Grammar (TROG-2) Version 2*. London: Psychological Corporation, 2003.

Breedin SD, Saffran EM, and Schwartz MF. Semantic factors in verb retrieval: An effect of complexity. *Brain and Language*, 63: 1–31, 1998.

Please cite this article in press as: Fraser KC, et al., Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2012.12.006

Brysbaert M and New B. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4): 977–990, 2009.

Chen M and Zechner K. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2011: p. 722–731. HLT '11.

Cheung H and Kemper S. Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics*, 13(01): 53–76, 1992.

Clark HH and Fox Tree JE. Using uh and um in spontaneous speaking. *Cognition*, 84(1): 73–111, 2002.

Cohen J, Asarnow R, Sabb F, Bilder R, Bookheimer S, Knowlton B, et al. Decoding continuous variables from neuroimaging data: Basic and clinical applications. *Frontiers in Neuroscience*, 5(75): 1–12, 2011.

Cotelli M, Borroni B, Manenti R, Alberici A, Calabria M, Agosti C, et al. Action and object naming in frontotemporal dementia, progressive supranuclear palsy, and corticobasal degeneration. *Neuropsychology*, 20(5): 558–565, 2006.

Croot K, Patterson K, and Hodges JR. Single word production in nonfluent progressive aphasia. *Brain and Language*, 61(2): 226–273, 1998.

Dunn LM and Dunn LM. *Peabody Picture Vocabulary Test*. 3rd ed. Circle Pines, Minnesota: American Guidance Service, 1997.

Folstein MF, Folstein SE, and McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12: 189–198, 1975.

Funnel E. W.L.P.: A case for the modularity of language function and dementia. In Code C, Wallesch CW, Joanette Y, and Lecours AR (Eds), *Classic Cases in Neuropsychology*. Hove, UK: Psychology Press, 1996.

Garrard P and Forsyth R. Abnormal discourse in semantic dementia: A data-driven approach. *Neurocase*, 16(6): 520–528, 2010.

Gilhooly K and Logie R. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods*, 12: 395–427, 1980.

Gorno-Tempini ML, Dronkers NF, Rankin KP, Ogar JM, Phengrasamy L, Rosen HJ, et al. Cognition and anatomy in three variants of primary progressive aphasia. *Annals of Neurology*, 55(3): 335–346, 2004.

Gorno-Tempini ML, Hillis AE, Weintraub S, Kertesz A, Mendez M, Cappa SF, et al. Classification of primary progressive aphasia and its variants. *Neurology*, 76: 1006–1014, 2011.

Graham NL, Patterson K, and Hodges JR. When more yields less: Speaking and writing deficits in nonfluent progressive aphasia. *Neurocase*, 10(2): 141–155, 2004.

Haenlein M and Kaplan AM. A beginner's guide to partial least squares analysis. *Understanding Statistics*, 3(4): 283–297, 2004.

Hall M, Frank E, Holmes G, Pfahringer B, Reutmann P, and Witten IH. The WEKA data mining software: An update. *ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations*, 2(1): 10–18, 2009.

Harciarek M and Kertesz A. Primary progressive aphasias and their contribution to the contemporary knowledge about the brain-language relationship. *Neuropsychology Review*, 21: 271–287, 2011.

Hillis AE, Oh S, and Ken L. Deterioration of naming nouns versus verbs in primary progressive aphasia. *Annals of Neurology*, 55(2): 268–275, 2004.

Hoerl AE and Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1): 55–67, 1970.

Jarrold W, Peintner B, Yeh E, Krasnow R, Javitz H, and Swan G. Language analytics for assessing brain health: Cognitive impairment, depression and pre-symptomatic Alzheimer's disease. In Yao Y, Sun R, Poggio T, Liu J, Zhong N, and Huang J (Eds), *Brain Informatics*. Lecture Notes in Computer Science. Springer Berlin/Heidelberg, 2010: 299–307.

Jurica PJ, Leitten CL, and Mattis S. *Dementia Rating Scale-2*. Lutz, FL: Psychological Assessment Resources, Inc, 2001.

Kaplan E, Goodglass H, and Weintraub S. *Boston Naming Test*. 2nd ed. Philadelphia: Lippincott Williams & Wilkins, 2001.

Kavé G, Leonard C, Cupit J, and Rochon E. Structurally well-formed narrative production in the face of severe conceptual deterioration: A longitudinal case study of a woman with semantic dementia. *Journal of Neurolinguistics*, 20(2): 161–177, 2007.

Klein D and Manning CD. Accurate unlexicalized parsing. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics: 423–430.

Knibb JA, Woollams AM, Hodges JR, and Patterson K. Making sense of progressive non-fluent aphasia: An analysis of conversational speech. *Brain*, 132(10): 2734–2746, 2009.

Kvalheim OM. Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots. *Journal of Chemometrics*, 24(7–8): 496–504, 2010.

Lambon Ralph MA, Graham KS, Ellis AW, and Hodges JR. Naming in semantic dementia – What matters? *Neuropsychologia*, 36(8): 775–784, 1998.

Li H. *Partial Least Squares-discriminant Analysis and Variable Selection for High Dimensional Data (Matlab Package)*, 2011.

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

Manning C. *Part-of-speech Tagging from 97% to 100%: Is It Time for Some Linguistics? Computational Linguistics and Intelligent Text Processing*, 171–189, 2011.

Manning CD, Raghavan P, and Schütze H. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Manning CD and Schütze H. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

Meteyard L and Patterson K. The relation between content and structure in language production: An analysis of speech errors in semantic dementia. *Brain and Language*, 110(3): 121–134, 2009.

Ng AY and Jordan MI. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In: Advances in Neural Information Processing Systems (NIPS) 14, vol. 2; 2002: 841–848.

Pakhomov SV, Smith GE, Chacon D, Feliciano Y, Graff-Radford N, Caselli R, et al. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and Behavioral Neurology*, 23: 165–177, 2010.

Patterson K and MacDonald MC. Sweet nothings: Narrative speech in semantic dementia. In Andrews S (Ed), *From Inkmarks to Ideas: Current Issues in Lexical Processing*. Hove, UK: Psychology Press, 2006.

Peintner B, Jarrold W, Vergyri D, Richey C, Tempini MLG, and Ogar J. Learning diagnostic models using speech and language measures. In: Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE: 4648–4651.

Raven JC. *Coloured Progressive Matrices Sets A, AB, B*. London: H. K. Lewis, 1962.

Rey A. L'examen psychologique dans les cas d'encephalopathie traumatique. *Archives de Psychologie*, 28: 286–340, 1941.

Roark B, Mitchell M, Hosom JP, Hollingshead K, and Kaye J. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7): 2081–2090, 2011.

Please cite this article in press as: Fraser KC, et al., Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2012.12.006

Saffran EM, Berndt RS, and Schwartz MF. The quantitative analysis of agrammatic production: Procedure and data. *Brain and Language*, 37: 440–479, 1989.

Sajjadi SA, Patterson K, Tomek M, and Nestor PJ. Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. *Aphasiology*, 26(6): 847–866, 2012.

Sampson G. Depth in English grammar. *Journal of Linguistics*, 33: 131–151, 1997.

Simons JS, Graham KS, Galton CJ, Patterson K, and Hodges JR. Semantic knowledge and episodic memory for faces in semantic dementia. *Neuropsychology*, 15(1): 101–114, 2001.

Singh V, Chertkow H, Lerch JP, Evans AC, Dorr AE, and Kabani NJ. Spatial patterns of cortical thinning in mild cognitive impairment and Alzheimer's disease. *Brain*, 129(11): 2885–2893, 2006.

Stadthagen-Gonzalez H and Davis CJ. The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, 38(4): 598–605, 2006.

Stopford CL, Snowden JS, Thompson JC, and Neary D. Variability in cognitive presentation of Alzheimer's disease. *Cortex*, 44(2): 185–195, 2008.

Stromswold K, Caplan D, Alpert N, and Rauch S. Localization of syntactic comprehension by positron emission tomography. *Brain and Language*, 52(3): 452–473, 1996.

Taler V and Phillips NA. Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. *Journal of Clinical and Experimental Psychology*, 30(5): 501–556, 2008.

Tanaka-Ishii K and Terada H. *Word Familiarity And Frequency*. Studia Linguistica, 2011. p. 96–116.

Thompson C, Ballard K, Tait M, Weintraub S, and Mesulam M. Patterns of language decline in non-fluent primary progressive aphasia. *Aphasiology*, 11: 297–321, 1997.

Thompson CK, Cho S, Hsu CJ, Wieneke C, Rademaker A, Weitner BB, et al. Dissociations between fluency and agrammatism in primary progressive aphasia. *Aphasiology*, 26(1): 20–43, 2012.

Tipping ME. Sparse Bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research*, 1: 211–244, 2001.

Toutanova K, Klein D, Manning C, and Singer Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 2003, p. 252–259.

Warrington EK and James M. *The Visual Object and Space Perception Battery*. Bury St Edmunds: Thames Valley Test Company, 1991.

Wilson SM, Henry ML, Besbris M, Ogar JM, Dronkers NF, Jarrold W, et al. Connected speech production in three variants of primary progressive aphasia. *Brain*, 133: 2069–2088, 2010.

Woollams AM, Cooper-Pye E, Hodges JR, and Patterson K. Anomia: A doubly typical signature of semantic dementia. *Neuropsychologia*, 46(10): 2503–2514, 2008.

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

Please cite this article in press as: Fraser KC, et al., Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2012.12.006