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