CORTEX 55 (2014) 122–129

Available online at www.sciencedirect.com

SciVerse ScienceDirect

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

Special issue: Research report

Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse

Peter Garrard [a,*], *Vassiliki Rentoumi* [a], *Benno Gesierich* [b], *Bruce Miller* [b] *and Maria Luisa Gorno-Tempini* [b]

[a] *Stroke and Dementia Research Centre, St George's, University of London, Cranmer Terrace, London SW17 ORE, UK*
[b] *UCSF Memory and Aging Center, Sandler Neurosciences Center, 675 Nelson Rising Lane, Suite 190, San Francisco, CA, USA*

ARTICLE INFO

Article history:
Received 15 January 2013
Reviewed 22 February 2013
Revised 12 March 2013
Accepted 15 May 2013
Published online 14 June 2013

Keywords:
Semantic dementia
Discourse
Laterality
Machine learning
Information gain

ABSTRACT

Advances in automatic text classification have been necessitated by the rapid increase in the availability of digital documents. Machine learning (ML) algorithms can 'learn' from data: for instance a ML system can be trained on a set of features derived from written texts belonging to known categories, and learn to distinguish between them. Such a trained system can then be used to classify unseen texts. In this paper, we explore the potential of the technique to classify transcribed speech samples along clinical dimensions, using vocabulary data alone. We report the accuracy with which two related ML algorithms [naive Bayes Gaussian (NBG) and naive Bayes multinomial (NBM)] categorized picture descriptions produced by: 32 semantic dementia (SD) patients versus 10 healthy, age-matched controls; and SD patients with left- (n = 21) versus right-predominant (n = 11) patterns of temporal lobe atrophy. We used information gain (IG) to identify the vocabulary features that were most informative to each of these two distinctions.

In the SD versus control classification task, both algorithms achieved accuracies of greater than 90%. In the right- versus left-temporal lobe predominant classification, NBM achieved a high level of accuracy (88%), but this was achieved by both NBM and NBG when the features used in the training set were restricted to those with high values of IG. The most informative features for the patient versus control task were low frequency content words, generic terms and components of metanarrative statements. For the right versus left task the number of informative lexical features was too small to support any specific inferences. An enriched feature set, including values derived from Quantitative Production Analysis (QPA) may shed further light on this little understood distinction.

1. Introduction

Rapid growth in the availability of digital documents, such as web pages, blogs and emails, has highlighted the importance of methods for automatic text classification. Imposing order on unstructured collections of texts facilitates storage, search, browsing and future re-use. Current approaches to text classification rely on machine learning (ML) techniques, which build automatic text classifiers by learning the characteristics of the categories of interest (such as topic or genre) from a set

* *Corresponding author.*
E-mail addresses: pgarrard@sgul.ac.uk, peter.garrard@gmail.com (P. Garrard).
0010-9452/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cortex.2013.05.008

of pre-classified documents. The many practical applications of ML-based text classification include authorship attribution, in which a text of unknown origin is attributed to one of a set of candidate authors. In some instances the approach has achieved accuracy comparable to that of human experts (Sebastiani, 2002). The potential of ML in the classification of language samples generated by different clinical populations has been less extensively investigated, though the problem of distinguishing among samples obtained from patients with different clinical syndromes is clearly analogous to that of authorship attribution.

Semantic dementia (SD) is a progressive neurodegenerative syndrome characterized by the relatively isolated degradation of the semantic component of long term declarative memory (Hodges et al., 1992; Snowden et al., 1992; Tulving, 1972; Warrington, 1975). Disruption of a system so central to language inevitably impacts on the production of discourse, giving rise to a pattern of spontaneous speech that is fluent, phonologically and grammatically correct, and makes use of a high frequency, often generic vocabulary (e.g., 'thing', 'bit' or 'stuff') (Bird and Lambon Ralph, 2000; Hodges et al., 1992; Meteyard and Patterson, 2009).

Studies of magnetic resonance (MR) imaging of the brains of patients with SD, at individual and group level, have consistently identified bilateral, asymmetric temporal lobe atrophy (Galton et al., 2001; Garrard and Hodges, 2000; Gorno-Tempini et al., 2004; Mummery et al., 2000) in which the left hemisphere is often (though not always see, e.g., Evans et al., 1995 and Thompson et al., 2003) the more profoundly affected. Comparisons between the subgroup with predominantly left- and predominantly right-sided atrophy (referred to henceforth as 'L > R' and 'R > L' respectively) have identified a number of features that appear to characterize the latter. These differences have largely emphasized the social and emotional impairment that frequently accompanies SD, the absence of insight (Chan et al., 2009; Perry et al., 2001; Thompson et al., 2003), and differences in the type of semantic knowledge disrupted (person specific rather than general) (Evans et al., 1995; Gentileschi et al., 2001; Josephs et al., 2008; Joubert et al., 2006).

In a recent study, Wilson et al. (2010) used Quantitative Production Analysis (QPA) (Saffran et al., 1989) to analyse transcripts of discourse samples obtained from patients with primary progressive aphasia (PPA) (Gorno-Tempini et al., 2011), including 25 with a diagnosis of SD. The SD transcripts differed most markedly from those of controls and other PPA syndromes on a lexical dimension, with striking differences identified in the increased used of pronouns, verbs, and high frequency nouns. Deviations from normal performance on selected aspects of syntactic structure and complexity were least marked in the SD subtype.

Studies examining the differential effects of L > R and R > L temporal lobe atrophy on language production are fewer in number: using measures derived from voxel based morphometry (VBM) (Good et al., 2001), Mummery et al. (2000) found that measures of semantic performance selectively correlated with atrophy in left-temporal lobe structures; Lambon Ralph et al. (2001) found that, in R > L cases, anomia and single word comprehension tended to deteriorate *pari passu*, while a L > R group showed disproportionately severe anomia. The same study also identified differences in the types of error made on a naming test by the two patient groups: the responses of L > R patients were more likely to be circumlocutory or omitted altogether, while patients in the R > L group committed more coordinate errors (such as production of 'goat' in response to a picture of a horse). To date, however, there have been no systematic comparisons of the effects of R > L versus L > R temporal lobe atrophy on the production of connected discourse in SD.

Garrard and Forsyth (2010) hypothesized that the language of SD patients and controls would be distinguishable on the basis of the lexical frequency data from transcripts of connected speech. They used principal components analysis (PCA) to identify two latent variables in a high dimensional 'discourse space', the values of which distinguished between connected speech samples obtained from patients with SD and those produced by controls. The study found that the vocabulary used by SD patients differed from those of controls along at least two major dimensions. The region of this two-dimensional 'discourse space' occupied by the control transcripts was characterized by the use of specific content-bearing terms and the deployment of a variety of grammatical function words. In contrast, the patient transcripts were associated with values correlating with use of the pronouns 'HE' and 'SHE', generic terms such as 'SOMETHING', the deictic words 'HERE' and 'THERE', the light verb 'DOING', and components of the phrase 'I DO NOT KNOW'. The two dimensions therefore captured not only the lexical-semantic deficit, but also the syntactically simplified character of SD discourse that has emerged from a number of manual analyses (Benedet and Patterson, 2006; Patterson and Macdonald, 2006).

In the current paper, we broaden this methodological approach by applying and comparing the performance of two (Bayesian) ML methods, in the classification of spoken discourse samples produced by SD patients and controls. In addition, we apply the same methods to the problem of distinguishing between speech samples produced by SD patients with L > R and R > L patterns of temporal lobe atrophy.

2. Materials and methods

The data for the study consisted of transcribed samples of connected speech, and structural MR imaging obtained from 32 patients meeting diagnostic and imaging criteria for SD (Gorno-Tempini et al., 2011; Hodges et al., 1992) and from 10 age matched, cognitively normal controls (NC). All participants were fluent English speakers, recruited through the Memory and Aging Center at UCSF after giving written informed consent. The study was approved by the institutional review board.

Within the SD group were 21 individuals classified on clinical and imaging grounds (see Section 2.2) as showing a L > R pattern of temporal lobe asymmetry, and 11 with the R > L pattern. The demographic characteristics of the three participant groups, as well as the performance of the SD groups on standardized tests of semantic memory, are summarized in Table 1. SD patients with a L > R pattern showed significantly greater impairment on naming and single word comprehension than the R > L subgroup, but the two SD

Table 1 – Demographic characteristics of the participant groups and standardized neuropsychological text scores obtained by the two groups of SD patients.

		SD L > R ($n = 21$)	SD R > L ($n = 11$)	Controls ($n = 10$)
Demographic characteristics	Mean [sd] age	62 [5.9]	64 [5.9]	68 [5.9]
	M:F	11:10	5:6	5:5
	n L-handed	3	1	0
Neuropsychological test (mean [sd])	BNT	3.0 [2.8]	7.6 [4.4][b]	n/a
	PPVT	6.1 [3.5]	9.7 [4.4][b]	n/a
	Fluency[a]	8.9 [.79]	9.4 [.81]	n/a
	Repetition (%)[a]	89.2 [10.7]	95.6 [4.3]	n/a
	PPTp (%)	39.1 [10.2]	42.8 [7.3]	n/a

BNT = Boston naming test; PPVT = Peabody Picture Vocabulary Test; PPTp = Pyramids and Palm Trees (picture condition).
a Subtests of the Western Aphasia Battery (Kertesz, 1982).
b $p < .05$ for the comparison between the two SD groups.

subgroups performed at equivalently low levels on tests of verbal fluency, repetition and visual associative semantics. There were no differences between the mean ages or gender distributions of the two patient subgroups, nor between the NC and SD groups.

2.1. *Connected speech samples*

Speech and language profiles of all participants had previously been examined using the Western Aphasia Battery (WAB) (Kertesz, 1982), which includes the elicitation of a sample of connected speech using the picnic picture description test. The examiner instructs participants to 'have a look at the picture, tell me what you see, and try to talk in sentences'. All test sessions were videotaped, and the audio components of these recordings were transcribed using standard English orthography.

Although the task is designed to elicit a monologue, in practice the interviewer is often heard supplying comments, prompts, or back-channel signals such as 'mmhmm' or 'yes' (Yngve, 1970), producing a dialogue that may at times deviate from the narrative of interest. Participants also produce paralinguistic elements such as 'um' and 'er' (generically coded as 'um'), exclamations such as 'well', or 'oh yeah', metanarrative elements (e.g., 'I can't remember the name'), inaudible words, and false starts. The transcriptions prepared for the present study incorporated these paralinguistic elements, but none of the interviewer's utterances.

2.2. *MR imaging*

Volumetric T1 images were acquired from all patients and controls on a 1.5T Siemens Magnetom VISION system equipped with a standard quadrature head coil, using a magnetization prepared rapid gradient echo (MPRAGE) sequence (164 coronal slices; slice thickness = 1.5 mm; FOV = 256 mm; matrix 256 × 256; voxel size 1.0 × q.5 × 1.0; TR = 10 msec; TE = 4 mm; flip angle = 15°). All images were compatible with the clinical diagnosis of SD: all showed focal atrophy of the inferolateral and polar aspect of one or both temporal lobes (Mummery et al., 2000). Images showing bilateral atrophy were further scrutinized to look for the presence and direction (R > L *vs* L > R) of any asymmetry. The classification was based on a consensus of senior clinicians (including MLG-T and BM), who took into account clinical information (such as the presence of greater deficits for famous people and social concepts, and reduced empathy) as well as structural imaging appearances.

3. Data analysis

All experiments were conducted using the Waikato Environment for Knowledge Analysis (WEKA) http://www.cs.waikato.ac.nz/ml/weka suite of ML software (Hall et al., 2009). Classifiers were trained using frequency unigrams (i.e., word types and their frequencies of occurrence in the groups of texts (corpora) to be classified). Transcripts were represented under the 'bag-of-words' assumption (i.e., absent any information relating to word order), which has been found to produce a robust solution in text classification by capturing word content- and frequency-specific differences that are relevant to the categories under investigation.

Each transcript was represented by a feature vector of the form:

$<w_1, x_1, \ldots, w_m, x_m>$ where $w_1,\ldots,w_m$ are the word types found in the set of transcribed texts (corpus) examined and $x_1 \ldots x_m$ denote the frequencies (i.e., the number of times the corresponding words occurred in a transcript). Each feature vector representation of a transcript is assigned to one of the categories, which is denoted by c_n. It should be noted that: (i) word types are derived from actual words rather than lemmatized forms; and (ii) the frequency assigned to each word type reflects its raw (rather than normalized) frequency of occurrence in each transcript.

For the first classification problem c_n took the values SD or NC, and for the second, L > R or R > L. We have a set of vector representations of transcripts already assigned to a class c_n (known as the 'training data') and we use these to train the classifier. Then we feed the trained classifier with a new set of unseen transcripts (the 'test data') and we expect it to map these new data to one of the classes of interest. For both classification tasks 10-fold cross validation was performed: data were divided into 10 subsets of equal size and the classification method trained 10 times, each time using a different combination of 9 subsets for training and the remaining subset for testing. The result is the average performance of the classifier on all 10 test sets.

3.1. ML classifiers

Two versions of the naive Bayes (NB) approach to ML classification were used. NB classifiers are implementations of Bayes' theorem, which concerns the degree to which the likelihood of a hypothesis being correct is contingent on previously unknown information. The term 'naive' derives from the fact that the classifier assumes the features it uses to classify texts to be conditionally independent given the class. Although the assumption of independence is rarely true, NB is reported to perform well even on complex tasks where it is clear that the strong independence assumptions are false (Russell and Norvig, 1998).

More formally, to calculate the probability of observing a feature vector $\vec{x}$ comprising features x_1 to x_m, given a class c under the NB assumption, the following holds:

$$p(\vec{x}|c) = p(x_1, \ldots, x_n|c) = \prod_{i=1}^{m} p(x_i|c). \quad (1)$$

In order for NB to be used as a classifier for a new transcript $\vec{x}$ it is easier to work with the posterior probability (i.e., that the hypothesis is correct given some new information):

$$p(c|\vec{x}) = p(c|x_1, \ldots x_n) \propto p(c)p(x_1|c)\ldots p(x_n|c). \quad (2)$$

or

$$p(c|\vec{x}) \propto p(c)\cdot p(\vec{x}|c)$$

where $p(c)$ refers to the prior probability that a transcript belongs to class c, which according to the maximum likelihood estimate is simply the ratio of the number of transcripts belonging to the particular class to the overall number of transcripts. The prior probability for the class SD is therefore (32/42) = .76, and for the class NC (10/42) = .24.

The NB classifier computes the class of each transcript by finding the one that maximizes the value of $p(c)\cdot p(\vec{x}|c)$, using the Bayes probabilistic model (equation 2), together with the maximum *a posteriori* (MAP) decision rule. Thus, NB classifies a transcript to a class using the classification function:

$$classify(\vec{x}) = argmax_c p(c) \prod_{i=1}^{m} p(x_i|c) \quad (3)$$

where c is one of the possible classes; $argmax_c$ indicates the class with the highest value for the function that follows it; $p(c)$ is the prior probability assigned to a given class; and $p(x_i|c)$ is the probability that the word feature with the value x_i belongs to a transcript of class c.

NB classifiers have a number of variants, which calculate $p(x_i|c)$ in subtly different ways (McCallum and Nigam, 1998). We now introduce the two versions used in the present study: Naive Bayes Gaussian (NBG) and Naive Bayes Multinomial (NBM). We describe how each version computes $p(\vec{x}|c)$, and how the values are exploited in the classification process.

3.1.1. NBG

The value of the probability $p(\vec{x}|c)$ is obtained under the assumption that the features are normally distributed across the transcripts in the corpus. The class for each transcript is therefore computed using the formula:

$$p(\vec{x}|c) = \prod_{i=1}^{m} g(x_i;\ \mu_{i,c},\ \sigma_{i,c}). \quad (4)$$

where $g(x_i..)$ is the normal distribution for each feature in each category c, μ is the mean, and σ is the standard deviation of these distributions (Metsis et al., 2006). By analogy with equations (3) and (4), we obtain the following classification function:

$$classify(\vec{x}) = argmax_c p(c) \prod_{i=1}^{m} g(x_i;\ \mu_{i,c},\ \sigma_{i,c}). \quad (5)$$

3.1.2. NBM

For this classifier, a multinomial distribution is assumed for each of the features. Such a feature distribution is assumed to work well with data that can easily be turned into frequencies, such as word counts in a text. If x_i is the frequency of a word w_i in a transcript d, then the probability of a transcript given its class is obtained from the formula:

$$p(\vec{x}|c) = p(|d|)\cdot|d|!\cdot \prod_{i=1}^{m} \frac{p(w_i|c)^{x_i}}{x_i!}. \quad (6)$$

and the classification function used to assign each transcript to a class from:

$$classify(\vec{x}) = argmax_c p(|d|)\cdot|d|!\cdot \prod_{i=1}^{m} \frac{p(w_i|c)^{x_i}}{x_1!}. \quad (7)$$

3.2. Feature selection algorithm

Not all features are of equal relevance in a classification problem, and identification of those making the largest contribution can improve classification performance as well as providing insights into the differences between the two document sets. To establish the relevance of individual unigram features to the classification methods described above, we used the Information Gain algorithm (Mitchell, 1997). This algorithm creates a decision tree to establish the purity of the groups into which each feature divides the to-be-classified instances. The information gain (IG) associated with a feature is defined as the difference in entropy (or randomness) of the sample before and after the creation of subsamples on the basis of that feature. For a binary classification into groups A and B, entropy (H) over a sample (T) is formally defined as:

$$H(T) = -P(A)\log(P(A)) - P(B)\log(P(B)). \quad (8)$$

where $P(A)$ and $P(B)$ are the probability density functions for A and B. The IG for a feature X in a sample T is:

$$IG(T, X) = H(T) - H(T|X). \quad (9)$$

which becomes:

$$IG(T,\ X) = H(T) - \sum_{v \in values(X)} \left(\frac{|T_v|}{|T|}\right) H(T_v). \quad (10)$$

where values (X) is the set of possible values for feature X and T_v is the subset of T in which feature X has value v. A feature whose values all belong to only one of the two categories would have an entropy of 0, giving the feature a very high value of IG, while one whose values always belong to both categories would have an entropy value closer to 1, and a lower IG value.

4. Results

4.1. Characteristics of the text corpus

The corpus of SD and NC transcripts contained 6381 individual word tokens and 744 unique word types. The SD corpus contained 4781 word tokens and 643 types. There were no differences between the mean word counts of the SD [150.7 (sd 71.9)] versus NC [155.8 (sd 54.1)] [$t(40) = .21$, $p > .05$], or of the L > R [149.2 (sd 62)] versus R > L [153.5 (sd 77.2)] transcripts [$t(30) = .15$, $p > .05$]; or between the mean type:token ratios within these groups [SD *vs* controls: $t(40) = .53$, $p > .05$; L > R *vs* R > L: $t(30) = .24$, $p > .05$].

4.2. Discriminating features

The features in the SD versus NC and L > R versus R > L distinctions that were associated with positive values of IG are listed in Table 2 (for the SD *vs* NC problem) and Table 3 (for L > R *vs* R > L), with the selected features ranked in order of IG.

Note that features are considered as relevant to the distinction rather than to one or other of the two classes. As an indicator of the relative importance of a feature to one class or the other, we also show the frequency of occurrence of each feature in the transcripts belonging to the two classes of interest. Consistent with earlier studies of discourse (Bird and Lambon Ralph, 2000; Garrard and Forsyth, 2010), the selected features that occur more often in NC transcripts consist of content words that appear in relatively low written and spoken frequency in published corpora (e.g., 'picnic', 'blanket', 'sailing' and 'pail'), while those occurring more often in SD scripts include generic terms (e.g., 'thing', 'something') and components of metanarrative statements (e.g., 'you', 'know' and 'remember'). The features found to be relevant to the L > R versus R > L distinction were fewer in number, but generally more indicative of one category over the other, from which it would be predicted that using selected features would result in a greater improvement in the classifiers' performance on this distinction. It is noted that six out of the seven content words in the list occur more frequently in the R > L transcripts suggesting a greater degree of semantic impairment in the L > R group, though metanarrative elements (such as 'am' and 'sure') are also more frequent in this group.

Finally, it should be noted that, as mentioned in 3.2, values of IG depend on how 'easy' it is for the algorithm to reach a result (i.e., in as few steps as possible in the decision tree). As an example, the word 'dog' appears about three times more often in SD transcripts, but does not have a high IG value. This is because the frequency values of the word can belong to either category, requiring the algorithm to go further down the decision tree before a concrete result can be obtained, increasing the entropy of the feature, and reducing its chances of being selected as a discriminative feature and thus assigned a high IG value. For IG it is important how distinctly distributed are the frequencies of the words in the two classes: the more different the frequency values of each word associated with the two classes, the more important it becomes for IG.

For the diagnostic classification, 43 word features (6% of the total) were selected out of the 744 unique words comprising the original feature set, while in the laterality classification only 14 words (2.2%) were selected from the original 643. These findings suggest that only a subset of features is likely to contribute to either of the two classification tasks, and raises the possibility that reduction of the dimensionality problem achieved using a feature selection approach could enhance the performance of classification algorithms. The dimensionality problem can be caused by large or redundant data sets in many classification tasks and can have an effect on the competence of classifiers to make efficient generalizations on unseen data. Classification accuracy will therefore be reported for analyses based on the entire set of features, and on the restricted sets.

Table 2 – Frequencies in transcripts of each class of words associated with positive values of information gain for the semantic dementia versus control classification task.

Feature	IG	Frequency	
		SD	NC
Picnic	.46	6	17
Kite	.34	10	12
Blanket	.32	4	13
Even	.32	1	0
Dock	.30	1	2
This	.29	47	0
With	.29	26	30
A	.27	287	169
Pail	.22	0	5
The	.21	228	127
Two	.21	4	9
People	.21	21	19
Flying	.20	19	14
Up	.20	23	0
Radio	.19	7	8
Playing	.18	5	10
Not	.17	25	0
Something	.17	24	0
Shore	.16	0	7
Flagpole	.16	0	3
Well	.16	29	1
Sailing	.15	2	5
Basket	.15	3	6
By	.15	4	6
Woman	.15	7	11
You	.15	53	3
Thing	.15	21	0
Flag	.15	13	11
Waving	.15	5	6
Know	.14	50	1
Down	.14	13	0
Off	.12	3	5
Remember	.12	16	0
Taken	.10	0	2
Apparently	.10	0	2
Behind	.10	0	2
Running	.10	0	2
Picnicking	.10	0	2
Relaxed	.10	0	3
Blowing	.10	0	3
Seem	.10	0	2
Lawn	.10	0	3
First	.07	3	0

Table 3 – Frequencies in transcripts of each class of words associated with positive values of information gain for the semantic dementia patients showing predominant left (L > R) versus predominant right (R > L) hemisphere atrophy.

Feature	IG	Frequency	
		L > R	R > L
Is	.22	54	78
Am	.16	1	7
Beach	.16	1	7
Blanket	.16	0	4
Flying	.16	8	11
Ok	.16	0	3
Picture	.16	0	3
Sailboat	.16	3	10
Sure	.16	0	4
Uh	.16	0	5
Now	.13	8	0
Drinking	.11	5	0
Fluid	.10	0	2

4.3. *Classification accuracy*

Accuracy of classification was measured by determining each model's ability correctly to assign a classification of SD versus NC, or L > R versus R > L, to each transcript, compared with a baseline condition in which all transcripts were assigned to the larger of the two classes. Results are expressed as the proportion of instances correctly classified. Accuracies achieved by each classification model were compared statistically to the baseline condition and to one another.

4.3.1. *SD versus controls*

Table 4 shows the confusion matrix for the SD versus NC classification solutions of the two classifiers using the complete feature set (i.e., all word types) (top half of the table), and those features that were selected by the IG algorithm (bottom half), together with their accuracy scores. Both classifiers show a significantly higher level of accuracy than that of the baseline condition (.76) ($p < .01$ for all comparisons) using both the full set of unigram features and the IG selected subset. Using the restricted feature set, the performance of NBG is marginally improved, while that of NBM is unchanged, though the former comparison failed to reach statistical significance.

4.3.2. *L > R versus R > L*

Table 5 displays the confusion matrices for the L > R versus R > L classification solutions of the two models, together with their accuracies, for both the full and the IG selected feature sets. When the full feature set was used, NBM yielded significantly greater accuracy ($p < .01$) than that of the baseline condition (.66), while NBG resulted in more classification errors (of both types) and accuracy that was not statistically different from baseline. Separate *t*-tests indicated that the accuracy achieved by NBM was significantly greater than that of NBG ($p < .05$). The NBG approach performed with greater accuracy when the reduced feature set was used, a difference that reached statistical significance ($p < .05$), while the accuracy of NBM was unaffected.

The significant improvement in the performance of the NBG classifier when it is presented with the restricted feature set is consistent with previous evidence of the adverse influence of redundant or irrelevant features on its performance (John et al., 1994; Rantanamahatana and Gunopoulos, 2003).

5. Discussion

The analyses reported in this paper support the idea that clinically relevant distinctions can be derived from samples of discourse using nothing more than the lexical frequency data inherent in the transcripts, as previously proposed by Garrard and Forsyth (2010).

The present study applied an analytical approach derived from the ML literature to a fresh set of transcripts, obtained from a larger group of patients using a different speech generation task. The average length of the speech samples subjected to the new analyses was also greater than those used in the 2010 study (150 *vs* 93.1 for SD and 153 *vs* 111.4 for controls), though there were no differences in length or simple lexical variety (type to token ratio) between any of the groups that were compared with one another. Moreover, to the task of separating SD from control discourse samples the present study added a novel discrimination between SD patients with distinct patterns of asymmetry in the degree of atrophy evident in the right and left-temporal lobes. Behavioural and

Table 4 – Classification accuracy achieved by the two machine learning algorithms on the SD versus NC categorization task, using the full lexical feature set (top) and only those features with positive information gain values (bottom).

			Naïve Bayes Gaussian		Naïve Bayes multinomial	
		Classification	SD	NC	SD	NC
All features	Category	SD ($n = 32$)	32	0	29	3
		NC ($n = 10$)	1	9	0	10
		Accuracy	.98		.93	
Selected features	Category	SD ($n = 32$)	32	0	29	3
		NC ($n = 10$)	0	10	0	10
		Accuracy	1		.93	

Table 5 – Classification accuracy achieved by two machine learning algorithms on the L > R versus R > L SD group categorization task, using the full lexical feature set (top) and only those features with positive information gain values (bottom).

			Naïve Bayes Gaussian		Naïve Bayes multinomial	
		Classification	L > R	R > L	L > R	R > L
All features	Category	L > R ($n = 21$)	17	4	21	0
		R > L ($n = 11$)	6	5	4	7
		Accuracy	.68		.88	
Selected features	Category	L > R ($n = 21$)	20	1	20	1
		R > L ($n = 11$)	3	8	3	8
		Accuracy	.88		.88	

neuropsychological differences between SD patients with these complementary patterns of atrophy have been extensively documented (Evans et al., 1995; Lambon Ralph et al., 2001; Thompson et al., 2003) but there had been no attempts to date to compare the characteristics of the connected speech produced by the two variants.

A variety of different approaches can be used to create ML classifiers (see Korde and Mahender (2012) for a review), but we chose to focus on the naive Bayes (NB) approach. Apart from a desire to minimize complexity in the presentation of methods and results, the motivation was that the specific models we adopted are particularly well suited to classifications involving word frequency data (Eyheramendy et al., 2003), though this does not mean that we consider other ML classifiers not to be worthy of investigation. Indeed, we intend to apply other models, such as support vector machines (Joachims, 1998), and structured models that can exploit word order and thus overcome the limitations of the 'bag-of-words' assumption, to these and similar clinical datasets.

To summarize our findings: using two variants (Gaussian (NBG) and multivariate (NBM)) of the naive Bayes approach, we found classifiers that could distinguish between SD patients and controls and between L > R and R > L SD patients with a high degree of accuracy. We also found that accuracy was influenced, to a variable extent, by the amount of redundancy (as indexed by low values of IG) in the feature set that was used for the classification. Both classifiers performed well in at least one of the two classification tasks, confirming that the general method, and the choice of unigram frequencies as feature data, were appropriate to the problem of classifying transcripts from these clinical groups. The differences in performance found between the two classifiers was, in most cases, not supported by statistical tests. The exception was the case of the L > R versus R > L classification task using the full feature set, in which the superior performance of NBM over NBG was significant. As described in Methods, NBM uses a multinomial distribution to describe the feature vector, while NBG uses a Gaussian distribution, of which the former better simulates the real distribution of the feature vector. Finally, our results showed how training the classifiers with a restricted set of automatically selected features can lead to increased accuracy, supporting the idea that reducing the amount of information available to a classifier can help it to achieve accurate distinctions between classes of interest. At a more practical level, a reduced set of feature can help to reduce the learning and running times of the classifiers – an important advantage when working with large data sets. The IG algorithm proved to be an effective way of achieving such feature reduction for the present purposes.

The fact that the categories of interest to this study were associated either with different vocabularies or different frequencies of certain usages can provide insights into how the language characteristics of these groups differ from one another. We used the IG algorithm to identify the set of words that contributed most to the two distinctions. In keeping with the findings of previous studies (Garrard and Forsyth, 2010; Bird and Lambon Ralph, 2000), the SD versus NC classification relied heavily on the presence of words that were overrepresented in one class or the other: generic and deictic terms (e.g., 'something' and 'this') and markers of metanarrative utterances (e.g., 'know' and 'remember'), were almost exclusively associated with the SD group; while a number of low frequency content-bearing words (e.g., 'shore', 'pail' and 'lawn') only occurred in NC descriptions. Note, however, that some words with high values of IG (e.g., 'kite' and 'people') are features of both groups of transcripts in roughly equal numbers, indicating that higher order attributes, such as words that typically occur in the same transcript, are of high discriminant value.

It is less easy to 'caricature' the speech typical of the L > R and R > L subgroups, due to the small number of features that achieved high values of IG. The relative overrepresentation of content-bearing words (e.g., 'sailboat', 'blanket' and 'beach') in the R > L transcripts would, however, be consistent with the idea that this clinical variant is associated with a milder semantic impairment than L > R (as was demonstrated by the comparative performances on more conventional neuropsychological tests). The differential frequencies of usages such as 'uh', 'ok' and 'sure' may, however, mark more subtle qualitative difference in style of delivery that would be more difficult to detect and quantify using standard instruments. Further exploration of this possibility in a ML framework, using additional observations derived from QPA (which include sentence and utterance length, the proportional occurrences of words belonging to individual grammatical classes, pauses, distortions, and other indicators of disordered phonology, and fluency) to enrich the feature set, will help to characterize this difference more precisely. Future studies should also compare the performance of other classifiers, explore the performance of the approach in languages other than English, and consider features based only on the presence or absence (rather than the frequency) of each word type in a transcript.

In conclusion, by applying an approach that was developed for and has proved useful in tasks related to digital document classification and authorship attribution to clinical language data, we have shown how clinically relevant distinctions can be supported using data derived entirely from transcripts of connected speech. Real-world clinical value would follow if the approach could be shown to be sensitive to more common clinical distinctions [such as between the connected speech of patients with early Alzheimer's disease, vascular cognitive impairment, and controls (Ahmed et al., 2012)]. Studies of language data acquired from members of these groups are currently in progress.

REFERENCES

Ahmed S, de Jager CA, Haigh AM, and Garrard P. Logopenic aphasia in Alzheimer's disease: Clinical variant or clinical feature? *Journal of Neurology Neurosurgery and Psychiatry*, 83: 1056–1062, 2012.

Benedet M and Patterson K. 'Non-semantic' aspects of language in semantic dementia: As normal as they're said to be? *Neurocase*, 12: 15–26, 2006.

Bird HE and Lambon Ralph MA. The rise and rall of frequency and imageability: noun and verb production in semantic dementia. *Brain Lang*, 73: 17–49, 2000.

Chan D, Anderson V, Pijnenburg Y, Whitwell J, Barnes J, Scahill R, et al. The clinical profile of right temporal lobe atrophy. *Brain*, 132: 1287–1298, 2009.

Evans JJ, Heggs AJ, Antoun N, and Hodges JR. Progressive prosopagnosia associated with selective right temporal-lobe atrophy – A new syndrome. *Brain*, 118: 1–13, 1995.
Eyheramendy S, Lewis DD, and Madigan D. On the Naive Bayes model for text categorization. In Bishop CM and Frey BJ (Eds), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, 2003*: 332–339.
Galton CJ, Patterson K, Graham K, Lambon-Ralph MA, Williams G, Antoun N, et al. Differing patterns of temporal atrophy in Alzheimer's disease and semantic dementia. *Neurology*, 57: 216–225, 2001.
Garrard P and Forsyth R. Abnormal discourse in semantic dementia: A data-driven approach. *Neurocase*, 16: 520–528, 2010.
Garrard P and Hodges JR. Semantic dementia: clinical, radiological and pathological perspectives. *Journal of Neurology*, 247: 409–422, 2000.
Gentileschi V, Sperber S, and Spinnler H. Crossmodal agnosia for familiar people as a consequence of right infero polar temporal atrophy. *Cognitive Neuropsychology*, 18: 439–463, 2001.
Good CD, Johnsrude IS, Ashburner J, Henson RN, Friston KJ, and Frackowiak RS. A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage*, 14: 21–36, 2001.
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: 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.
Hall M, Frank E, Holmes G, Prahringer 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: 10–18, 2009.
Hodges JR, Patterson K, Oxbury S, and Funnell E. Semantic dementia. Progressive fluent aphasia with temporal lobe atrophy. *Brain*, 115(Pt 6): 1783–1806, 1992.
Joachims T. Text caegorisation with support vector machines: Learning with many relevant features. *European Conference on Machine Learning*, 137–142, 1998.
John GH, Kohavi R, and Pfleger K. Irrelevant features and the subset selection problem. In Cohen WW and Hirsh H (Eds), *International Conference on Machine Learning New Brunswick, NJ*, 1994.
Josephs KA, Whitwell JL, Vemuri P, Senjem ML, Boeve BF, Knopman DS, et al. The anatomic correlate of prosopagnosia in semantic dementia. *Neurology*, 71: 1628–1633, 2008.
Joubert S, Felician O, Barbeau E, Ranjeva JP, Christophe M, Didic M, et al. The right temporal lobe variant of frontotemporal dementia: Cognitive and neuroanatomical profile of three patients. *Journal of Neurology*, 253: 1447–1458, 2006.
Kertesz A. *Western Aphasia Battery*. New York: Grune and Stratton, 1982.
Korde V and Mahender CN. Text classification and classifiers: A survey. *International Journal of Artificial Intelligence and Applications*, 3: 85–99, 2012.
Lambon Ralph MA, McClelland JL, Patterson K, Galton CJ, and Hodges JR. No right to speak? The relationship between object naming and semantic impairment: Neuropsychological abstract evidence and a computational model. *Journal of Cognitive Neuroscience*, 13: 341–356, 2001.
McCallum A and Nigam K. A comparison of event models for naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*. Madison, Wisconsin, 1998.
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: 121–134, 2009.
Metsis V, Androutsopoulos I, and Paliouras G. Spam filtering with naive bayes – Which naive bayes? Conference on Email and Anti-spam (CEAS). vol. 17. Mountain View, CA, 2006: 29–69.
Mitchell T. *Machine Learning*. McGraw-Hill, 1997.
Mummery CJ, Patterson K, Price CJ, Ashburner J, Frackowiak RS, and Hodges JR. A voxel-based morphometry study of semantic dementia: Relationship between temporal lobe atrophy and semantic memory. *Annals of Neurology*, 47: 36–45, 2000.
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: Psychology Press, 2006.
Perry RJ, Rosen HR, Kramer JH, Beer JS, Levenson RL, and Miller BL. Hemispheric dominance for emotions, empathy and social behaviour: evidence from right and left handers with frontotemporal dementia. *Neurocase*, 7: 145–160, 2001.
Rantanamahatana C and Gunopoulos D. Feature selection for the naive Bayesian classifier using decision trees. *Applied Artificial Intelligence*, 17: 475–487, 2003.
Russell S and Norvig P. *Artificial Intelligence: A Modern Approach*. Pearson, 1998.
Saffran EM, Berndt RS, and Schwartz MF. The quantitative-analysis of agrammatic production – Procedure and data. *Brain and Language*, 37: 440–479, 1989.
Sebastiani F. Machine learning in automated text categorization. *ACM Computing Surveys*, 34: 1–47, 2002.
Snowden JS, Neary D, Mann DMA, Goulding PJ, and Testa HJ. Progressive language disorder due to Lobar atrophy. *Annals of Neurology*, 31: 174–183, 1992.
Thompson SA, Patterson K, and Hodges JR. Left/right asymmetry of atrophy in semantic dementia – Behavioral-cognitive implications. *Neurology*, 61: 1196–1203, 2003.
Tulving E. Episodic and semantic memory. In Tulving E and Donaldson W (Eds), *Organization of Memory*. New York: Academic Press, Inc., 1972.
Warrington EK. Selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology*, 27: 635–657, 1975.
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.
Yngve V. On getting a word in edgewise. In *Sixth Regional Meeting of the Chicago Linguistic Society*, 1970.