Journal of Alzheimer's Disease 42 (2014) S3–S17
DOI 10.3233/JAD-140555
IOS Press

Features and Machine Learning Classification of Connected Speech Samples from Patients with Autopsy Proven Alzheimer's Disease with and without Additional Vascular Pathology

Vassiliki Rentoumi[a], Ladan Raoufian[a], Samrah Ahmed[b], Celeste A. de Jager[c] and Peter Garrard[a,*]
[a]*Neuroscience Research Centre, Cardiovascular and Cell Sciences Research Institute, St. George's, University of London, London, UK*
[b]*Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK*
[c]*Institute of Ageing in Africa, Division of Geriatric Medicine, Department of Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa*

Accepted 3 July 2014

Abstract. Mixed vascular and Alzheimer-type dementia and pure Alzheimer's disease are both associated with changes in spoken language. These changes have, however, seldom been subjected to systematic comparison. In the present study, we analyzed language samples obtained during the course of a longitudinal clinical study from patients in whom one or other pathology was verified at post mortem. The aims of the study were twofold: first, to confirm the presence of differences in language produced by members of the two groups using quantitative methods of evaluation; and secondly to ascertain the most informative sources of variation between the groups. We adopted a computational approach to evaluate digitized transcripts of connected speech along a range of language-related dimensions. We then used machine learning text classification to assign the samples to one of the two pathological groups on the basis of these features. The classifiers' accuracies were tested using simple lexical features, syntactic features, and more complex statistical and information theory characteristics. Maximum accuracy was achieved when word occurrences and frequencies alone were used. Features based on syntactic and lexical complexity yielded lower discrimination scores, but all combinations of features showed significantly better performance than a baseline condition in which every transcript was assigned randomly to one of the two classes. The classification results illustrate the word content specific differences in the spoken language of the two groups. In addition, those with mixed pathology were found to exhibit a marked reduction in lexical variation and complexity compared to their pure AD counterparts.

Keywords: Alzheimer's disease, computational methods, diagnosis, language, machine learning, vascular dementia

INTRODUCTION

Engaging in spoken communication is a complex, multidimensional skill that draws on a wide variety of cognitive domains, including semantic memory, syntactic knowledge, and phonological abilities, and is consequently dependent on a large and widely distributed network of cortical and subcortical brain

*Correspondence to: Dr. Peter Garrard, Neuroscience Research Centre, Cardiovascular and Cell Sciences Research Institute, St. George's University of London, Cranmer Terrace, London SW17 0RE, UK. Tel.: +44 208 725 5117; Fax: +44 208 725 2950; E-mail: pgarrard@sgul.ac.uk.

ISSN 1387-2877/14/$27.50

regions. It is not surprising, therefore, that degenerative conditions, such as Alzheimer's disease (AD), in which the pathological lesion is diffuse, are commonly associated with deficits across a range of subcomponents of linguistic competence [1–3].

Although language deterioration may precede clinical recognition of AD [4], it typically manifests in the early stages of the disease in the form of word finding difficulty, and is often accompanied by a mild to moderate degree of anomia related in large part to the disintegration of semantic memory [5]. Analyses of spoken discourse at early disease stages, based on the results of standardized assessment schedules, have confirmed the presence of semantic impairment [3] as well as more subtle abnormalities, such as an increased rate of lexical errors and a reduction in syntactic complexity [6].

Impairment of spoken language is also a feature of vascular cognitive impairment (VCI) and vascular dementia (VaD), though evidence that the linguistic profile differs from that of AD is equivocal. Some comparisons of language related test performance in patients with clinical diagnoses of AD and VaD, based on batteries of speech and language tests, have suggested a greater variety of language deficits in the latter group [7]. In particular, patients with VaD typically exhibit difficulties on measures of verbal fluency and motor aspects of speech, probably due to frontal executive deficits, which are less pronounced in AD [7, 8]. Other studies [7, 9] report a greater reduction in syntactic complexity in the speech of VaD patients, with conciseness well maintained in both groups. In contrast, Kontiola et al. [10], found that VaD patients had most difficulty with basic language abilities such as recognition of words, naming, and repetition, while AD patients had more difficulties with understanding and producing complex grammatical structures. Comparisons based on narrative speech, however, have revealed less distinct patterns of impairment [5].

The inconsistency of these comparative studies may be due to variations in the methods of assessment used (some involved naming tasks while others involved elicitation of connected speech), but almost certainly reflects, in addition, the frequent coexistence of the two pathologies, which limits their definitive distinction at the clinical level. Finally, at the time of clinical presentation, patients may have declined on cognitive and linguistic measures beyond the stage at which their profiles can be readily distinguished [2].

The approach adopted in the present study overcame these limitations by introducing a number of modifications to previous methods. In the first place, we used retrospective language data consisting of samples of connected speech taken from archived testing sessions that had been conducted on participants in a longitudinal study of ageing and dementia. In the study in question (the Oxford Project to Investigate Memory and Ageing (OPTIMA) [11], serial clinical evaluation was performed annually until death, and in more than 80% of cases post mortem examination was carried out. As a result, we were able to ensure not only that language sampling took place at a uniform stage of clinical progression, but also that the disease groups could be defined according to the gold standard of pathological information obtained at postmortem, avoiding the uncertainty and circularity of relying on clinical diagnosis. Secondly, in order to mimic the practical scenario in which clinical differentiation would most commonly be required, we divided the patients according to the degree of vascular pathology that was present at postmortem in addition to, rather than instead of, plaque and tangle pathology.

Analysis of the language samples was conducted using automatic feature selection and machine learning (ML) classification algorithms [12] to identify key distinctive characteristics and use them to maximize the correct diagnostic classification by pathological group. These methods take account of the multidimensional and probabilistic nature of pattern recognition, but have only recently appeared in the field of dementia diagnosis: Fraser et al. in [13] applied a ML approach using features that could be automatically extracted from digital samples of connected speech, to distinguish among different subtypes of primary progressive aphasia. Garrard et al. [15] applied ML classification techniques to distinguish discourse samples produced by patients with semantic dementia from those of normal controls, but based solely on the lexical features of the two sets of transcripts. Here, we use a combination of automatically extracted pure lexical, and more complex linguistic (e.g., syntactic and textual), characteristics to define the features of interest, and test their predictive value with ML classifiers.

MATERIALS AND METHODS

Study sample

The study sample consisted of transcripts of connected speech, recorded in the course of regular (six to twelve monthly) interval assessments of participants in OPTIMA—a longitudinal study of aging and dementia in a cohort of elderly, community living volunteers. In many individuals who entered the study with normal

cognition, longitudinal assessment included conversion to and progression through successive stages of dementia. Interval assessment included neuroimaging, blood and cerebrospinal fluid sampling, physical examination and, of particular interest to this study, neuropsychological evaluation. At entry into the study, all participants were invited to consent to post mortem examination. Participants included in the current study had enrolled in OPTIMA between 1988 and 2008. All were either classified as either cognitively normal or as meeting criteria for mild cognitive impairment (MCI) at study entry. Diagnoses were regularly updated during follow-up assessments, with an increasing proportion of patients reclassified as probable AD over time. Data from 36 subjects were used for the present study, all of whom had been followed up until death, and in all of whom a pathological diagnosis of AD had been made at post mortem. The appearances documented in the pathologist's report allowed categorisation of participants into two equally sized groups: 1) Mixed AD (AD_m): plaque and tangle pathology together with cerebrovascular disease that was described in the pathologist's report as being moderate to severe; 2) Pure AD (AD_p): plaque and tangle pathology with absent or minimal cerebrovascular change.

Participants in both groups fell within the mild to moderate range of cognitive impairment (as defined by a: Mini-Mental State Examination (MMSE) scoring range 10–25) at the time of language sampling, and were matched as closely as possible for age, gender distribution, and years of education.

Table 1 shows group comparisons of demographics and cognitive performance based on assessment using the CAMCOG instrument. CAMCOG, which forms part of the Cambridge Mental Disorders of the Elderly Examination (CAMDEX) [15] interview, comprises eight major subtests that correspond to different areas of cognitive function, as well as a derived MMSE score [16]. CAMCOG covers a broader range of cognitive domains than the MMSE, assessing orientation, language, memory, praxis, attention, abstract thinking, perception, and calculation, and was administered at every assessment date for each participant, with sessions tape-recorded and archived for future use.

As Table 1 shows, there were no significant differences in age or gender distribution between the two groups, nor in the number of years spent in full time education. While there was no significant difference in performance on the MMSE between the AD groups, Tukey *post hoc t*-tests confirmed significant differences on 7 out of 11 CAMCOG subtests between the AD_p and AD_m group. Of particular note, participants with AD_m performed significantly worse (p-values ≤ 0.012) than ADp in: expression, remote memory, learning memory, attention, calculation, abstract thinking, and perception. Orientation, comprehension, recent memory, and praxis were not significantly different ($p > 0.05$).

Table 1
AD_p and AD_m groups' comparison of demographics and cognitive performance. Comparisons AD_p and AD_m patient groups computed using Tukey's *post-hoc* tests and Chi-square for comparison of gender ratio; Maximum scores given in parentheses

	AD_p ($n = 18$)		AD_m ($n = 18$)		
	Mean	SD	Mean	SD	
Demographics					
Age (y)	74.3	9.13	75.1	7.34	n.s
Education (y)	13.1	3.1	11.7	3.3	n.s
Gender (male: female)	12 : 6		6 : 12		n.s
MMSE (30)	21.1	3.6	18.6	4.8	n.s
CAMCOG scores					
Total (107)	75.6	8.3	58.2	17.1	****
Orientation (10)	6.6	2.3	6.3	1.8	n.s
Comprehension (9)	7.9	1.2	7.6	1.3	n.s.
Expression (21)	16.2	1.9	13.7	3.7	*
Remote memory (6)	4.4	1.3	2.8	1.5	***
Recent memory (4)	2.5	1.0	1.8	1.3	n.s
Learning memory (17)	7.0	3.6	4.1	3.0	**
Attention (7)	5.1	1.9	3.1	2.6	**
Praxis (12)	9.4	1.6	8.1	0.8	n.s
Calculation (2)	1.7	0.6	0.9	0.8	****
Abstract thinking (8)	5.7	1.9	3.2	2.9	**
Perception (11)	9.1	1.4	6.6	2.7	****

Means (M) and standard deviations (SD) of demographic data. Cognitive performance n.s. = not significant. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$. MMSE, Mini-Mental State Examination.

Language data

The data for the present study consisted of a total of 36 transcribed samples (Supplementary Material) of connected speech obtained from the two groups. These transcripts were obtained using the 'Cookie theft' picture description test, which forms part of the Boston Diagnostic Aphasia Examination [17], and is also a component of the CAMCOG. All test sessions were transcribed using standard English orthography, following the conventions described in Garrard et al. [18].

Transcript analysis

The aim of the analysis was to classify every transcript correctly with respect to its subgroup (AD_p versus AD_m) of origin. The analysis took place in three consecutive stages, with the output of the first two stages each providing input for the next.

In the first stage, features were extracted from all transcripts to create a set of values corresponding to distinct features of each text. We considered features as belonging to three broad categories: i) purely lexical features, which consisted of word types and their frequencies (token counts) found within each transcript; ii) syntactic complexity; and iii) textual features, which included indices of lexical variation [19, 20], and measures derived from information theory. Features were obtained using Lu's L2 Syntactic Complexity Analyzer [22], and Keyplex (a textual analysis program whose output consists of values for a range of statistical and lexical measures). In addition, a compression ratio feature was computed using the zlib library of the Python programming language.

The second stage of the analysis consisted of feature selection, in which those features that differentiated the transcripts of participants with AD_p from those with AD_m were identified. The final stage was ML classification, in which the selected features were used to 'train' the ML classifier to assign transcripts to one of the two groups.

The ML, feature selection, and text classification stages of the methodology were implemented using the Waikato Environment for Knowledge Analysis (WEKA; http://www.cs.waikato.ac.nz/ml/weka) [23]. Methodology for all stages is described in detail in the following three sections.

Feature extraction

A detailed list of automatically extracted features and their definitions can be found in Table 2.

The first two features, which form the purely lexical category, are frequency and binary unigrams (i.e., word types and their frequencies of occurrence or presence/absence values) derived from each transcript. A total of 529 frequency/binary unigram features were extracted. Such 'pure lexical' information has proved useful in discriminating discourse samples produced by patients with semantic dementia from those of normal controls [4]. In that study, as here, the transcripts were represented under the 'bag-of-words' assumption (i.e., without any information relating to word order), which has been found to produce a robust solution in text classification by capturing word content-related and frequency-related differences that are relevant to the categories under investigation.

The choice to employ purely word frequency and word presence/absence related features is also supported by the observation that the poorer language ability seen in patients with VaD is associated with less lexical variety and decreased complexity [24]. We might also expect certain words that appear in the AD_m vocabulary to be of lower frequency than those in the AD_p vocabulary [5].

The next 14 features belong to the syntactic complexity category and were calculated using Lu's L2 Syntactic Complexity Analyzer [22]. Syntactic complexity is defined by Ortega in [21] as the range, and degree of sophistication, of forms that surface in language production. Lu used these features to analyze the syntactic complexity of college-level English essays from Chinese students, and the tool has also been employed in an analysis of narrative speech transcripts produced by patients with different subtypes of primary progressive aphasia [13] and right versus left temporal lobe predominant semantic dementia [14].

The next 7 features belong to the textual category: the first 5 (features 17–21) are relevant to lexical complexity and variation and the last two to information theory measures. Using lexical variation measures [19, 20] we were able automatically to measure the vocabulary range in each patient group as displayed in their transcripts. In particular, Honoré's R [22] takes into account the probability that the participant will reuse a given word type in the text rather than using a new one. This is calculated as the ratio of the words occurring once only in the vocabulary (hapax legomena), to the total number of distinct words. Thus, for a given text length, the value of R is higher when there are more hapax legomena and thus less repetition. We also used a related measure, Simpson's D [23], which gives the probability that two words that have been randomly selected from the text are the same. This measure quantifies the rate of word repetition in samples; lower values of D indicating less repetitive texts. Similarly, in the dis legomena over vocabulary measure (i.e., words that occur twice divided by the total number of different words appearing in the transcript), higher values of the measure imply a poorer vocabulary. On the other hand, higher rates of hapax legomena and pair-hapax legomena (i.e., once used token-pairs) imply a more varied language use.

The first information theory feature uses Shannon entropy (H) [24] computed in transcripts. In information theory, H is equivalent to the amount of information (measured in bits) that is added when the value of a previously unknown variable is obtained. The entropy of a random variable equates to its unpredictability. Shannon showed how information content [25] could also be measured in written language, empirically determining the accuracy with which a reader could predict the identity of sequentially

Table 2
Definitions of features extracted from AD_p and AD_m data sets. *Values of these variables represented the mean of the values computed across sequential blocks of 10 words within each speech transcript

	Features category	Features definition
	Purely Lexical	
1	Binary Unigrams	Presence/Absence of word types
2	Frequency Unigrams	Frequency of word types
	Syntactic Complexity	
3	MLC (Mean Length of Clause)	number of words/number of clause (clause: a structure consisting of at least a subject and a finite verb)
4	MLS (Mean Length of Sentence)	number of words/number of sentences
5	MLT (Mean Length of T-Unit)	number of words/number of T-units (T-Unit: one main clause plus any subordinate clause or non clausal structure that is attached to or embedded in it)
6	C/S (Sentence complexity ratio)	number of clauses/number of sentences
7	C/T (T-unit complexity ratio)	number of clauses/number of T-units
8	CT/T (Complex T-unit ratio)	number of complex T-units/number of T-units complex T-unit: a T-unit that contains a dependent clause
9	DC/C (Dependent clause ratio)	number of dependent clauses/number of clauses dependent clauses: clause which could not form a sentence on its own
10	DC/T (Dependent clauses per T-unit)	number of dependent clauses/number of T-units
11	CP/C (Coordinate phrases per clause)	number of coordinate phrases/number of clauses coordinate phrase: a phrase immediately before a coordinating conjunction (e.g. and, or)
12	CP/T (Coordinate phrases per T-unit)	number of coordinate phrases/number of T-units
13	T/S (Sentence coordination ratio)	number of T-units/number of sentences
14	CN/C (Complex nominals per clause)	number of complex nominal/number of clauses (Complex nominals: they comprise (i) nouns plus adjective, possessive, prepositional phrase, relative clause, participle, or appositive, (ii) nominal clauses, and (iii) gerunds and infinitives in subject position)
15	CN/T (Complex nominals per T-unit)	number of complex nominals/number of T-units
16	VP/T (Verb phrases per T-unit)	number of verb phrases/number of T-units
	Textual	
17	Honore R (R)	$R = 100 \log(N)/(1-V_1/V)$ N = text length V_1 = number of hapax legomena (once-used tokens), V = number of different words (types)
18	*Mean of hapax legomena	number of hapax legomena (once-used tokens)/text block of 10 words
19	*Mean of pair-hapax legomena	mean number pair-hapaxlegomena (i.e. of once-used token-pairs)/text block of 10 words
20	*Mean of Simpson's diversity index (D)	$D = \sum_{i=1}^{v} i(i-1)\frac{V(i,N)}{N(N-1)}$ N = number of word occurrences (i.e. tokens), V (i, N) = the number of types which occur i times in a sample of N tokens, and v the highest frequency of occurrence
21	Dis legomena over Vocabulary	number of dis legomena (tokens that are repeated twice)/number of different words (types)
22	*Shannon Entropy (H) Mean	H/text block of 10 words $H(X) = -\sum_{i=1}^{n} p(x_i)\log_b p(x_i)$ Minus is used because for values less than 1 logarithm is negative
23	Compression Ratio (CR)	Compressed Size of a file (zipped)/Uncompressed size of a file (unzipped)

revealed characters (including spaces) in a segment of text. Thus, H is related to the information content of language insofar as less patterned sequences (in this case, sentences) convey high entropy and are highly informative. On the other hand, a patterned and predictable sequence conveys low entropy, and is less informative.

The final measure used was another information theory feature, the compression ratio (CR) [26], which was employed to compute the repetitiveness in each patient's transcript. Compression is achieved by condensing a piece of data such that it takes up less space than it did originally, but still contains the same amount of information [27]. Compression on texts works by reducing the redundancy of a text by omitting words that are repeated, allowing a repetitive text to be compressed to a greater extent than a non-repetitive one. Compression ratio is the ratio of the size of the compressed text to its original (uncompressed) size. From the above, it follows that the notions of entropy and compression are tightly correlated. Intuitively, we expect a text with low entropy and low information content to have a high compression capacity, since its structure will be more patterned and predictable, while a text with higher entropy carrying more new information will have a smaller compression potential.

Both the lexical variation and the information theory features are included in the textual category, as vocabulary studies [28] manifest a strong inter-correlation among all these kinds of features. In practice, increased lexical richness is associated with an increased entropy, and thus with a greater degree of randomness and uncertainty, which would be expected to characterize a text with less repetitiveness and a lower compressibility. We would expect our experiments to confirm this assumption.

Since some textual category features such as hapax legomena, pair-hapax legomena, Simpson's diversity index (D) and entropy (H) are sensitive to document length we computed their mean values across sequential blocks of 10 words within each speech transcript. We adopted this convention in order to normalize the text length, as the transcript lengths vary from 26 to 164 words.

Feature selection

Not all features are of equal relevance in a classification problem, and identification of those with the highest impact can improve classification performance, as well as providing insights into between-group language differences. To establish the relevance of the features described above, we used a correlation-based feature selection approach (CFS) [29]. CFS is a filter approach that uses a correlation-based heuristic to determine the usefulness of features. According to this approach the unnecessary or redundant features are filtered out of the data.

CFS uses a search algorithm together with a function to evaluate the usefulness of feature subsets. Specifically, the heuristic approach employed by CFS measures the appropriateness of each feature for predicting a class label—in this case the AD_p and AD_m groups—together with the level of intercorrelation among them. The function used to compute these correlations is based on conditional entropy [30]. The resulting feature subsets contain features that are highly correlated with the class, yet uncorrelated with each other. These feature subsets are then evaluated using a ML classification approach (see below). CFS was selected because it is considered a good fit in cases where large numbers of features are employed [29].

Machine learning text classification

A ML classification approach was used to predict the group (AD_p or AD_m) to which each participant's transcript belonged, based solely on the features selected in the step described above. For each transcript, the corresponding values of the selected features formed a feature vector representation of the transcript. The ultimate goal of a ML text classification is to accept a feature vector of an unknown group and provide as output a group label for it (in this case, either AD_p or AD_m).

Representation of narrative speech transcripts by vectors

In order to perform classification exploiting the binary and the frequency unigram feature sets (see above) representing the transcripts, we consider that each transcript was represented by a feature vector of the form: $<w_1x_1, \ldots, w_mx_m>c_n$, where $w_1, \ldots, w_m$ are the word types[1] found in the set of transcribed texts examined, and $x_1 \ldots x_m$ denote the frequencies (i.e., the number of times the corresponding words occurred in a transcript) or the presence/absence of each word type in the transcript. Within the vector representation word types are used as labels for words for the sake of completeness in order to indicate that each word feature is assigned a value, word tokens themselves have not been used in the classification process. Each feature vector representation of a transcript is assigned to one of the groups, 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.

In using the remaining categories of syntactic complexity and textual features, the classification was performed using the following representation for each transcript: $<m_1x_1, \ldots, m_kx_k>c_n$, where $m_1, \ldots, m_k$ are the corresponding features extracted from the transcripts. $x_1, \ldots, x_k$ denote the numeric output value of each of these specific features computed on the transcript, and c_n indicates to which of the two groups AD_p, AD_m each transcript is assigned during the classification task.

In general, a ML classifier is provided with an input of a set of vector representations of transcripts already assigned to a group (the 'training data'). Then the trained classifier (i.e., classification model) is provided with a new set of unseen transcripts (the 'test data') and expected to map these new data to one of the groups of interest. In our case, the output categories would be either AD_p or AD_m. The effectiveness of each configu-

[1] In the vector representation the word types are used as labels in order to indicate the type to which each value belongs. That is needed as types with zero value are omitted. They are not used as features in the classification process.

ration of the classification algorithm is measured using a 3-fold cross validation procedure (described below).

Naive bayes classifier

The Naïve Bayes (NB) classifier is an implementation of Bayes' theorem. The term 'naïve' derives from the fact that the classifier assumes the features it uses to classify texts to be conditionally independent given the category. 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 [31].

The NB classifier learns estimates for the category-conditional probabilities and priors for each group (AD_p, AD_m) from the training data. At the classification stage, Bayes theorem is used to assign a feature vector representing a transcript to the class that maximizes the posterior probability (i.e., the more likely class) [31]. NB classifiers have a number of variants, which calculate the probability of a transcript belonging to a class in subtly different ways; among the most popular NB variants are the Naïve Bayes Gaussian (NBG) and the Naïve Bayes Multinomial (NBM) [32]. In NBG, the value of the probability that each transcript belongs to a group is obtained under the assumption that the features are normally distributed across the transcripts in the corpus. In NBM, a multinomial distribution is assumed for each of the features representing a transcript. In the current work we made use of NBG, as it was found to outperform NBM in the classification of transcripts from patients with semantic dementia [15].

Evaluation procedures

To evaluate both the feature selection and classification procedures, we adopted a three-fold cross-validation approach as follows. In each classification task, the transcripts used were randomly divided into three equally-sized, randomly selected, subsets. Two subsets were used to select the features, which were then applied on the same two subsets to train the classification algorithm. The remaining subset was used for testing the classification algorithm, based on the selected features. This process was repeated three times, using differently constituted test subsets at each iteration.

RESULTS

Feature extraction

An example output of feature extraction is shown in Table 3, which lists the lexical features extracted from the AD_m versus AD_p data set. Features are shown along with their mean values for each group of transcripts. In order to keep the list within manageable limits, we selected from the lexical features category those items whose frequency unigrams were associated with an average frequency greater than 0.1 in either of the two groups under comparison.

Evaluation of feature selection

The CFS feature selection method was tested under the three distinct feature representations (binary unigrams, frequency unigrams, and combined syntactic complexity and textual variation features)[2].

All features belonged to the categories detailed in Table 2.

Feature selection output

The output of each feature selection task comprises three selected feature subsets, one from each iteration of the evaluation procedure, which are tested against the remaining data subsets at each iteration. Table 4 shows the selected features for each feature selection task, as well as the mean values of the normalized frequency and standard deviation of each selected feature for each group of transcripts.

In the case of the combined syntactic complexity and textual variation features representation, none of the features were selected from the feature selection task, and consequently in Table 4 we report the differences in mean values from the whole set, across the three iterations.

Features selected from the pure lexical category

Table 4 shows three feature subsets ranging from nine to twelve words, corresponding to the binary and frequency feature representation. Each feature is accompanied by its mean value and standard deviation over the full data set.

In coarse-grained terms, the selected binary discriminative feature set indicates the words that have a tendency to appear in one group, while the selected word frequency features indicate the words that appear more frequently in one group over the other. The differences in mean values of the selected binary and frequency unigrams show that in most of the cases a selected word feature either appears only in the AD_p

[2]Preliminary experimentation with combined feature sets (i.e., frequency unigrams feature set together with the combined syntactic complexity and textual feature set) showed poor discrimination, presumably due to the introduction of noise, so the results are reported for individual feature sets.

Table 3
Word features (i.e., unigrams) extracted from the AD_p versus AD_m data set of transcripts assigned with their mean frequency of appearance in the transcripts of each group (i.e., in AD_p versus AD_m)

Frequency unigrams extracted from the AD_p versus AD_m language data set	Mean frequency of appearance		Frequency unigrams extracted from the AD_p versus AD_m language data set	Mean frequency of appearance	
	AD_p	AD_m		AD_p	AD_m
they	0.33	0.33	catch	0.06	0.00
not	0.33	0.22	look	0.17	0.11
now	0.11	0.06	will	0.11	0.00
drop	0.11	0.00	while	0.22	0.00
biscuits	0.16	0.06	apron	0.11	0.00
side	0.11	0.11	is	2.44	1.72
mean	0.05	0.11	it	1.11	0.94
doing	0.38	0.44	in	1.00	1.00
house	0.11	0.00	things	0.11	0.17
out	0.89	0.39	plates	0.11	0.17
washed	0.11	0.06	hand	0.50	0.11
looking	0.22	0.00	off	0.50	0.39
stupid	0.00	0.11	i	0.56	1.28
got	0.78	0.44	no	0.00	0.17
quite	0.17	0.06	well	0.61	0.72
put	0.11	0.11	mother	0.39	0.56
her	0.94	0.39	the	7.06	6.28
could	0.11	0.00	left	0.11	0.11
running	0.33	0.28	just	0.17	0.11
thing	0.17	0.11	when	0.17	0.00
think	0.28	0.17	cakes	0.44	0.33
one	0.56	0.17	yes	0.22	0.67
feet	0.17	0.00	cut	0.00	0.11
another	0.11	0.00	thinking	0.17	0.00
open	0.22	0.06	has	0.33	0.11
little	0.83	0.83	big	0.00	0.11
top	0.17	0.06	gonna	0.11	0.17
too	0.06	0.28	know	0.06	0.22
really	0.00	0.11	presume	0.00	0.17
curtains	0.17	0.11	bit	0.00	0.22
that	0.89	1.28	lady	0.44	0.17
shelf	0.22	0.00	knock	0.11	0.00
huh	0.00	0.11	like	0.22	0.17
tree	0.11	0.00	lost	0.00	0.11
and	4.44	4.17	because	0.22	0.11
say	0.00	0.17	some	0.17	0.06
have	0.33	0.17	back	0.33	0.17
she	1.56	1.44	standing	0.50	0.00
dishes	0.17	0.28	dear	0.11	0.06
take	0.11	0.00	saucer	0.00	0.11
which	0.72	0.17	for	0.39	0.17
though	0.11	0.06	asking	0.11	0.11
who	0.06	0.17	be	0.44	0.56
had	0.00	0.11	reaching	0.17	0.06
taps	0.11	0.22	by	0.22	0.06
busy	0.06	0.00	on	1.39	0.72
state	0.00	0.06	sister	0.28	0.06
should	0.06	0.17	ok	0.00	0.11
only	0.06	0.00	getting	0.56	0.22
going	0.28	0.22	of	1.56	0.50
pretty	0.06	0.00	or	0.11	0.39
do	0.11	0.11	into	0.11	0.06
his	0.61	0.28	socks	0.11	0.00
topple	0.11	0.00	right	0.00	0.22
get	0.33	0.17	there	1.06	1.28
drawn	0.06	0.11	was	1.06	1.28

Table 3
(*Continued*)

Frequency unigrams extracted from the AD_p versus AD_m language data set	Mean frequency of appearance		Frequency unigrams extracted from the AD_p versus AD_m language data set	Mean frequency of appearance	
	AD_p	AD_m		AD_p	AD_m
him	0.06	0.17	head	0.11	0.00
overflowing	0.44	0.11	himself	0.00	0.11
puddle	0.11	0.00	but	0.11	0.06
we	0.06	0.11	wiping	0.28	0.11
up	1.11	1.61	trying	0.11	0.11
see	0.11	0.28	with	0.44	0.28
are	0.37	0.59	he	0.28	0.11
said	0.06	0.17	whether	1.56	1.44
away	0.00	0.11	um	0.11	0.00
outside	0.17	0.06	cake	1.28	0.83
mum	0.00	0.11	an	0.33	0.22
probably	0.22	0.06	as	0.11	0.00
kitchen	0.22	0.28	at	0.22	0.17
taking	0.17	0.11	girl	0.11	0.06
against	0.06	0.00	again	0.61	1.00
forgotten	0.06	0.06	floor	0.11	0.00
presumably	0.11	0.17	tip	0.22	0.22
fallen	0.11	0.06	tin	0.11	0.00
been	0.28	0.06	you	0.22	0.17
wet	0.11	0.00	er	0.06	0.33
wise	0.06	0.06	helping	2.17	0.78
child	0.11	0.00	happening	0.00	0.17
plate	0.28	0.22	ooh	0.17	0.00
minute	0.00	0.11	box	0.00	0.11
so	0.05	0.27	window	0.11	0.00
mm	0.11	0.00	roof	0.50	0.11

transcripts, or if it appears in both groups, it has a higher frequency in the AD_p transcripts. This shows that the differences between the two groups, as far as the selected lexical features are concerned, are both quantitative and qualitative: i.e., that the two vocabularies contain different words, and that common words occur in different frequencies.

The majority of the selected binary and frequency features are content words (e.g., 'garden', 'hand', 'sister', 'feet'). Of these content words, those referring to concrete objects or events appear either exclusively, or with greater frequency, in the AD_p vocabulary. In contrast, the remaining word types, which are either verbal or adverbial forms (e.g., standing, right), refer to more abstract concepts and are not exclusive to one group or the other. It is notable that verb types included in the binary and frequency unigram subsets are morphologically complex (i.e., 'standing', 'getting', 'looking') and that they occur in higher frequency in the AD_p group.

Features selected from the syntactic complexity and textual categories

No specific features were selected during the feature selection step, so all the features were used. As CFS selects the features that are uncorrelated with other features, this was to be expected from the basis on which they are calculated (see above, 'Feature extraction'). As an example, a text that has a lot of repetition will consequently have low entropy and compression ratio. Additionally, the higher the repetition, the lower the entropy and the compression ratio, and vice versa.

Table 4, shows all the syntactic complexity and textual features accompanied by the mean value and standard deviation for each group. It is evident that the mean values of the majority of the syntactic complexity measures employed are higher in the AD_p transcripts. The large number of clauses per sentence, the large number of verb phrases per T-unit, and the increased length of sentences, which characterize the AD_p group indicate a more complex syntax. Complex syntax indicates preservation of expressive ability and less impaired language function. For instance, the following two passages (the first produced by an AD_p patient, and the second by an AD_m) are examples of low and high syntactic complexity, respectively:

(i) ''er it's the boy is um wanting to pinch some biscuits from a top shelf to do so he's standing

on a stool *which* is not a very stable thing to stand on *however* the girl has got her wits about her *because* she's got her hand ready to catch the biscuits *when* the boy has got hold of them providing *that* he doesn't land on his back and er knock himself out *in the meantime* the er

Table 4
Lexical features selected by the CFS, and (unselected) syntactic complexity and textual features for the AD_p versus AD_m groups' distinction for each of the 3 iterations in each feature selection task. Each feature is accompanied by the mean and standard deviation over the tested transcripts belonging to each group

Features category	Selected features	AD_p (Mean/SD)	AD_m (Mean/SD)
Binary Unigrams			
1st iteration	pinching	0.00 (0.16)	0.25 (0.43)
	window	0.42 (0.49)	0.00 (0.16)
	looking	0.25 (0.43)	0.00 (0.16)
	one	0.50 (0.5)	0.00 (0.16)
	up	0.67 (0.47)	1.00 (0.16)
	when	0.25 (0.43)	0.00 (0.16)
	has	0.33 (0.47)	0.00 (0.17)
	bit	0.00 (0.16)	0.25 (0.43)
	standing	0.50 (0.5)	0.00 (0.16)
	for	0.50 (0.5)	0.08 (0.27)
	you	0.00 (0.17)	0.25 (0.43)
2nd iteration	garden	0.25 (0.43)	0.00 (0.17)
	so	0.00 (0.17)	0.33 (0.47)
	feet	0.25 (0.43)	0.00 (0.17)
	is	1.00 (0.17)	0.75 (0.43)
	hand	0.42 (0.49)	0.00 (0.17)
	know	0.00 (0.17)	0.25 (0.43)
	standing	0.58 (0.49)	0.00 (0.17)
3rd iteration	pinching	0.00 (0.17)	0.25 (0.43)
	drying	0.50 (0.5)	0.08 (0.27)
	too	0.00 (0.17)	0.33 (0.47)
	no	0.00 (0.17)	0.25 (0.43)
	bit	0.00 (0.17)	0.25 (0.43)
	standing	0.41 (0.49)	0.00 (0.17)
	sister	0.33 (0.47)	0.00 (0.17)
	right	0.00 (0.17)	0.25 (0.43)
	er	0.83 (0.37)	0.25 (0.43)
Frequency Unigrams			
1st iteration	garden	0.25 (0.43)	0.00 (0.17)
	so	0.00 (0.17)	0.42 (0.64)
	out	1.00 (0.81)	0.33 (0.47)
	feet	0.25 (0.43)	0.00 (0.17)
	hand	0.58 (0.76)	0.00 (0.17)
	know	0.00 (0.17)	0.25 (0.43)
	standing	0.67 (0.62)	0.00 (0.17)
	getting	0.75 (1.08)	0.25 (0.43)
	of	1.90 (1.8)	0.40 (0.6)
2nd iteration	to	2.43 (1.99)	1.31 (1.33)
	pinching	0.00 (0.17)	0.25 (0.43)
	too	0.00 (0.17)	0.33 (0.47)
	no	0.00 (0.17)	0.25 (0.43)
	bit	0.00 (0.17)	0.25 (0.43)
	standing	0.50 (0.64)	0.00 (0.17)
	sister	0.33 (0.47)	0.00 (0.17)
	Er	2.50 (1.80)	0.60
3rd iteration	very	0.38 (0.62)	0.00 (0.17)
	two	0.30 (0.46)	0.00 (0.17)
	pinching	0.00 (0.17)	0.23 (0.42)
	when	0.23 (0.42)	0.00 (0.17)
	bit	0.00 (0.17)	0.23 (0.42)
	standing	0.46 (0.50)	0.00 (0.17)
	you	0.00 (0.17)	0.38 (0.62)

Table 4
(*Continued*)

Features category	Selected features	AD_p (Mean/SD)	AD_m (Mean/SD)
Combined Syntactic Complexity and Textual Features			
Average Values across the 3 iterations	MLC (Mean Length of Clause)	7.90 (2.83)	6.39 (1.57)
	MLS (Mean Length of Sentence)	85.65 (42.93)	52.12 (28.52)
	MLT (Mean Length of T-Unit)	65.91 (44.57)	47.46 (29.77)
	C/S (Sentence complexity ratio)	11.23 (5.04)	8.74 (4.49)
	C/T (T-unit complexity ratio)	8.50 (4.04)	7.70 (4.7)
	CT/T (Complex T-unit ratio)	0.77 (0.32)	0.71 (0.33)
	DC/C (Dependent clause ratio)	0.54 (0.23)	0.43 (0.16)
	DC/T (Dependent clauses per T-unit)	5.29 (3.5)	3.65 (2.5)
	CP/C (Coordinate phrases per clause)	0.16 (0.14)	0.15 (0.17)
	CP/T (Coordinate phrases per T-unit)	1.11 (0.6)	1.13 (0.5)
	T/S (Sentence coordination ratio)	1.37 (0.7)	1.24 (0.5)
	CN/C (Complex nominals per clause)	0.78 (0.30)	0.64 (0.22)
	CN/T (Complex nominals per T-unit)	6.83 (5.5)	4.87 (3.31)
	VP/T (Verb phrases per T-unit)	10.25 (4.69)	8.1 (5.1)
	Honore R (R)	2085.12 (488.82)	1784.02 (540.50)
	Mean of hapaxlegomena	8.83 (0.55)	8.47 (0.99)
	Mean of pair-hapaxlegomena	9.94 (0.11)	9.83 (0.35)
	Mean of Simpson's diversity index (D)	0.87 (0.00)	0.88 (0.01)
	Dislegomena over Vocabulary	0.13 (0.04)	0.16 (0.05)
	Shannon Entropy Mean (H)	0.990 (0.0)	0.99 (0.01)
	Compression Ratio (CR)	0.59 (0.05)	0.56 (0.05)

presumably the mother is getting on with the washing up"

(ii) "looks as if the boy's pinching cakes *and* the girl's washing up apparently oh there's a flood isn't there yes *and* the stool is tipping up"

The first example consists of a single, long sentence with numerous dependent clauses (introduced by the words in bold type). The frequent use of verb phrases indicates a complex syntax. By contrast the second sentence comprises a series coordinate phrases (i.e., phrases dominating a coordinating conjunction ('and', 'but', 'for', 'or', 'nor', 'yet', and 'so')). Increased complexity at the clausal level indicates greater eloquence, as the correct use of subordinate clauses is associated with greater linguistic ability, while, the higher phrasal density—such as the greater amount of coordination—entails a simpler mode of expression and reduced eloquence.

Moreover, the mean values of textual features imply that the AD_p vocabulary is more varied: values of Honoré's R and rates of hapax legomena and pair-hapax legomena over segments of equal size are all higher in the AD_p transcripts. From the mean value of the dis legomena over vocabulary, it appears that the AD_p vocabulary exhibits a smaller number of words that are used more than once, yielding a less redundant vocabulary. The slightly higher value of Shannon entropy (H) in the AD_p transcripts has the same implication. On the other hand, the lower mean value of compression ratio, which characterizes the AD_m transcripts, indicates that the latter are somewhat more repetitive [6].

Machine learning classification results

The ML classifier was also evaluated following the three-fold cross validation approach described above. In the current section, we present accuracy results for five ML classification tasks performed for the AD_p/AD_m distinction: three classification tasks based on the features selected as described above, and two further classification tasks using the entire frequency and binary feature sets.

Each classification task consisted of three sub-experiments, corresponding to the three test sets described above. In the first three tasks, these sub-experiments exploited the feature subsets from the three feature representations summarized in Table 4, with three separate feature subsets forming the input vectors for the NB classifier. The features subsets for the remaining two classification tasks were taken from the entire frequency and binary feature sets. The success of each sub-experiment is determined by the

Table 5
Confusion matrices and accuracy for Naïve Bayes (NB) using various feature sets for the distinction between the AD_p and AD_m groups. Accuracy comparisons (error threshold 5%), *(w/FS): With feature selection, *(w/o FS): Without feature selection, NB: Naïve bayes

	NB* with binary unigrams (w/FS)*		NB with binary unigrams (w/o FS)*		NB with frequency unigrams (w/FS)		NB with frequency unigrams (w/o FS)		NB with combined syntactic complexity & textual features		Baseline	
	AD_m	AD_p	AD_m	AD_p	AD_m	AD_p	AD_m	AD_p	AD_m	AD_p	AD_m	AD_p
AD_m	13	5	11	7	13	5	10	8	13	5	18	0
AD_p	7	11	8	10	4	14	6	12	8	10	18	0
Accuracy	66.67%		58.33%		75%		61.11%		63.89%		50%	

1) NB with Binary Unigrams (w/FS), NB with Binary Unigrams (w/o FS) VS Baseline (p-values: 0.01, 0.02), NB with Frequency Unigrams (w/FS), NB with Frequency Unigrams (w/o FS) VS Baseline (p-values: 0.001, 0.02), NB Combined Syntactic Complexity & Textual Features VS Baseline: p-value: 0.01. 2) NB with Binary Unigrams (w/FS) VS NB with Binary Unigrams (w/o FS): p-value: 0.02, NB with Frequency Unigrams (w/FS) VS NB with Frequency Unigrams (w/o FS), p-value: 0,03. 3) NB with Frequency Unigrams (w/FS) VS NB with Binary Unigrams (w/FS), NB with Binary Unigrams (w/o FS), NB with Frequency Unigrams (w/o FS), NB Combined Syntactic Complexity & Textual Features: (p-values: 0.02, 0.005, 0,01, 0.02).

classifier's accuracy, which is defined as the percentage of correctly classified transcripts. This outcome was compared statistically with a baseline condition in which all transcripts were assigned to one of the two classes at random. Since the number of samples was equal (18 transcripts) for each category, the baseline accuracy performance was considered as 50%. For the binary and frequency feature sets, classification performance with and without feature selection were also compared. Finally, comparison of the accuracy of the classifier across feature sets additionally allowed the relative effectiveness of each feature set to be compared. In all comparisons statistical significance was tested using paired t-tests against a 5% error threshold. Table 5 shows the confusion matrices and average accuracy of the classifier derived from the three subexperiments for each classification task, along with p-values for the comparisons of interest.

In all the classification tasks, NB produced a significantly higher level of accuracy than the baseline condition. For both frequency and binary feature sets, the classification tasks using selected features outperformed the tasks without feature selection. The results also indicate that the selected frequency feature set was more effective in distinguishing the two categories than the selected binary feature set, while the equivalent tasks without feature selection showed similar levels of accuracy.

The configuration with the highest accuracy (75%) was the frequency feature set after feature selection, suggesting that the contrast in the language of these groups is to a large extent due differences in the frequencies of certain words. Moreover, the significantly superior performances of combined feature set compared to baseline, supports the intuition that differences in these groups can also be located at the syntactic complexity and lexical variation levels.

DISCUSSION

Longitudinal cohort studies of patients with neurodegenerative conditions have given rise to a number of important insights linking cognitive performance during life with neuropathological appearances at postmortem. In the current study we aimed to detect the extent to which computational methods could accurately detect the most important sources of variation in connected language samples produced by patients with two distinct species of neuropathology, namely, those with changes of AD alone ('pure AD', referred to in this study as AD_p) and those with a combination of AD and vascular pathology ('mixed AD', referred to here as AD_m). We also aimed to determine the extent to which such automated methods could accurately distinguish between the two groups. While previous studies have found differences based on a combination of manually and automatically derived features of texts, and applied a ML approach to clear cut clinical distinctions (such as that those between fluent and non-fluent progressive aphasia [13]), the present study used only features that could be extracted automatically from digitized text samples, and focused on a more subtle (though no less clinically important) distinction.

Experimental results clearly supported the idea that the use of features ranging from purely lexical to more complex syntactic and textual, combined with a ML classification approach, could successfully distinguish between the two groups. The results also showed how the automatically extracted feature subsets were

distributed between the two patient groups. Concretely, information about the presence or absence of lexical features and the frequency of words appeared to be a valuable indicator of the vocabulary specific differences between the two groups. Classification results also suggested that a combination of syntactic complexity, information and textual variation was important to the same clinical distinction. These findings are consistent with earlier clinical observations of patients with AD_p and AD_m in the language domain, with reference to both the syntactic and lexical fluency levels [4, 5]. Finally, feature selection suggested that characteristics defined using an information theory approach is also important to this clinical distinction. These issues will be discussed in more detail in the following sections.

Feature selection

The selected binary and frequency feature subsets selected proved effective at distinguishing between the AD_p versus AD_m subsets when integrated with the classification tasks, while examination of the selected words in both the binary and frequency subsets (Table 4) yielded insights into the idiosyncrasies of the language of the two groups. Despite the fact that the selected word features in both subsets were few in number, they were selected from a larger total (529 words) and can be considered to be representative. The selections suggest that a range of content words contribute to the difference between the groups, though the mere presence (or absence) or number of content words in isolation is not adequate to the task of accurate classification. The power of CFS is that it indicates a minimum subset of features whose correlation with the groups under investigation makes them capable of discriminating between the groups. This cannot be achieved by manually performed descriptive approaches.

With respect to the textual and syntactic complexity features, the feature selection algorithm did not select any features so all the features were used. There are two possible explanations for this result: as the first is that the features reflect well established metrics whose values correlate in a predictable fashion (more repetition leading to lower entropy and higher compression ratio), and features that are predictive of one another (or most of the others in that case) are excluded from the features selected by CFS. An alternative explanation is that none of the features was highly correlated with the class, and are therefore not very important in distinguishing between the groups. Although intuition would seem to favor the former explanation, we acknowledge that formal comparison of CFS with other feature selection methods (which is beyond the scope of the present study) would be needed to exclude the latter.

The study introduced a series of features derived from information theory. When used in conjunction with syntactic complexity and textual features in the raw set, these features contributed to an efficient differentiation of the two groups. The differences in mean values between the two groups of the full feature sets from the syntactic complexity and textual domains (as shown in Table 4) indicate that in texts with a higher degree of syntactic complexity with numerous subordinate clauses (as in the case of the AD_p transcripts), less repetition is expected since the introduction of new lexical items is more probable in such "rich" structures. On the other hand our findings verified the intuitive expectation of a reduction in syntactic complexity, especially in the AD_m group, and are in agreement with previous attempts at investigating the effects of AD [33] on syntactic complexity. Similarly, the higher entropy in the AD_p group accords with the more varied vocabulary observed for this group [3], while verifying that increased lexical richness is largely associated with higher entropy [28], and thus with a greater degree of randomness and uncertainty. This association reveals an analogy between lexical richness and entropy, which is explained by considering that the bigger number of different words in a text, the more difficult to predict the occurrence of a given word.

Machine learning classification

The 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, a task in which ML has been shown to perform as accurately as human experts [34]. The potential of ML to classify the language of different clinical populations has been less extensively investigated, though the task is clearly an analogous one. In the present study, a ML classification approach exploiting a selected set of words with either presence/absence or frequency information, yielded good diagnostic accuracies of 67% and 75% respectively, which was superior to that of a combination of textual and syntactic complexity features. By implementing a threefold cross validation approach, we eliminated the possibility that the classifier over-fitted on the language idiosyncrasies of the training set, causing poor generalizability to unseen data. Instead, features were selected only from the training set (2/3 of the dataset), which

was different at each iteration, and were always tested through the application of the classification method on the test set (1/3 of the dataset).

The accuracy, as a result of the threefold cross validation across all the classification tasks performed, means that the classifier's performance was high in all the threefold classification tasks comprising the cross-validation approach. It is also important to note, however, that high accuracy can also be seen as a result of the effective combination of the appropriate ML algorithm with the most appropriate feature subset. The reason behind the choice of NB was its previously demonstrated good fit for text classification tasks [35]. Moreover, the NB classifier has been reported to perform better when it deals with non-redundant features [36]. The rationale behind the choice to employ a reduced set of features, was that it also helps to reduce the learning and running times of the classifiers—an important advantage when working with large data sets. The CFS algorithm proved, most of the time, to be an effective way of achieving such a feature reduction for the present purposes. The performance of the classifier, when integrating the combined syntactic complexity and textual feature sets achieved a significantly higher performance than that of the baseline condition. This argues in favor of the existence of systematic differences in syntactic complexity and lexical variation between the two groups. The accuracy achieved by ML classification of the distinctions between the two groups was indicated by its consistent outperforming of the baseline condition, but also by the levels of classification accuracy achieved (as high as 75%).

In summary, these findings, using a ML method for narrative transcript analysis, illustrated an important set of differences in the language features of postmortem confirmed AD_m and AD_p groups. They verified our initial hypothesis that the cognitive deficits of AD_m are greater than in AD_p, and evidenced by a greater deficit for the AD_m in both lexical variety and syntactic complexity in spoken language.

ACKNOWLEDGMENTS

We thank all study participants and their families, without whose long-term commitment, this study could not have occurred. Similarly, we wish to thank Professor A. David Smith, founder and former director of OPTIMA, who set up this longitudinal cohort study, and Professor Gordon Wilcock, Project Director, who kindly agreed to share the data with the authors. Dr. Richard Forsyth wrote the Keyplex programme. OPTIMA nurses and staff contributed to the data collection and neuropsychological assessment of participants. Yasmin Galbraith conducted preliminary experimentation related to the current study. This work was supported by a UK Medical Research Council grant to PG [G0801370].

All data were acquired after approval of the Oxford Project to Investigate Memory and Aging (OPTIMA) by the local ethics committee.

Authors' disclosures available online (http://www.j-alz.com/disclosures/view.php?id=2433).

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: http://dx.doi.org/10.3233/JAD-140555.

REFERENCES

[1] Ahmed S, Haigh A-MF, de Jager CA, Garrard P (2013) Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain* **136**, 3727-3737.

[2] Habash A, Guinn C, Kline D, Patterson L (2012) Language analysis of speakers with dementia of the Alzheimer's type. *Annals of the Master of Science in Computer Science* and *Information Systems at UNC Wilmington* V6 N1 Paper 11, http://csbapp.uncw.edu/data/mscsis/annalspaper.aspx?v=6&i=1&p=11

[3] Ahmed S, de Jager CA, Haigh A-M, Garrard P (2013) Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease. *Neuropsychology* **27**, 79-85.

[4] Garrard P, Maloney LM, Hodges JR, Patterson K (2005) The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain* **128**, 250-260.

[5] Vuorinen E, Laine M, Rinne J (2000) Common pattern of language impairment in vascular dementia and in Alzheimer disease. *Alzheimer Dis Assoc Disord* **14**, 81-86.

[6] Ahmed S, de Jager CA, Haigh A-MF, Garrard P (2012) Logopenic aphasia in Alzheimer's disease: Clinical variant or clinical feature? *J Neurol Neurosurg Psychiatry* **83**, 1056-1062.

[7] Powell AL, Cummings JL, Hill MA, Benson DF (1988) Speech and language alterations in multinfarct dementia. *Neurology* **38**, 717-717.

[8] Desmond DW (2004) The neuropsychology of vascular cognitive impairment: Is there a specific cognitive deficit? *J Neurol Sci* **226**, 3-7.

[9] Hier DB, Hagenlocker K, Shindler AG (1985) Language disintegration in dementia: Effects of etiology and severity. *Brain Lang* **25**, 117-133.

[10] Kontiola Pi, Laaksonen R, Sulkava R, Erkinjuntti T (1990) Pattern of language impairment is different in Alzheimer's disease and multi-infarct dementia. *Brain Lang* **38**, 364-383.

[11] Budge MM, de Jager C, Hogervorst E, Smith AD (2002) Total plasma homocysteine, age, systolic blood pressure, and

cognitive performance in older people. *J Am Geriatr Soc* **50**, 2014-2018.

[12] Michie D, Spiegelhalter DJ, Taylor CC (1994) *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, New York.

[13] Fraser KC, Meltzer JA, Graham NL, Leonard C, Hirst G, Black SE, Rochon E (2014) Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex* **55**, 43-60.

[14] Garrard P, Rentoumi V, Gesierich B, Miller B, Gorno-Tempini ML (2014) Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. *Cortex* **55**, 122-129.

[15] Roth M, Tym E, Mountjoy C, Huppert FA, Hendrie H, Verma S, Goddard R (1986) CAMDEX. A standardised instrument for the diagnosis of mental disorder in the elderly with special reference to the early detection of dementia. *B J Psychiatry* **149**, 698-709.

[16] Folstein MF, Folstein SE, McHugh PR (1975) "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* **12**, 189-198.

[17] Goodglass H, Kaplan E (1983) *The Assessment of Aphasia and Related Disorders*. Lea & Febiger, Philadelphia.

[18] Garrard P, Haigh A-M, de Jager C (2011) Techniques for transcribers: Assessing and improving consistency in transcripts of spoken language. *Lit Linguist Comput* **26**, 389-405.

[19] Duran P, Malvern D, Richards B, Chipere N (2004) Developmental trends in lexical diversity. *Appl Linguist* **25**, 220-242.

[20] Yu G (2010) Lexical diversity in writing and speaking task performances. *Appl Linguist* **31**, 236-259.

[21] Ortega L (2003) Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Appl Linguist* **24**, 492-518.

[22] Honoré A (1979) Some simple measures of richness of vocabulary. *Assoc Literary Linguist Comput Bull* **7**, 172-177.

[23] Simpson EH (1949) Measurement of diversity. *Nature* **163**, 688.

[24] Shannon CE (1951) Prediction and entropy of printed English. *Bell Syst Tech J* **30**, 50-64.

[25] Shannon CE (1959) Coding theorems for a discrete source with a fidelity criterion. *IRE Nat Conv Rec* **4**, 142-163.

[26] Salomon D, Motta G, Motta G (2010) *Handbook of Data Compression*, Springer, London.

[27] Behr F, Fossum V, Mitzenmacher M, Xiao D (2003) Estimating and comparing entropies across written natural languages using PPM compression. *Proceedings of the 2003 Data Compression Conference*, p. 416.

[28] Thoiron P (1986) Diversity index and entropy as measures of lexical richness. *Comput Hum* **20**, 197-202.

[29] Hall MA, Smith LA (1997) Feature subset selection: A correlation based filter approach. *International Conference on Neural Information Processing and Intelligent Information Systems*, 855-858.

[30] Press WH, Teukolsky SA, Vetterling WT, Flanneryt BP (2007) *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, New York.

[31] McCallum A, Nigam K (1998) A comparison of event models for naive Bayes text classification. In *AAAI/ICML98 Workshop on Learning for Text Categorization*, pp. 41-48.

[32] Metsis V, Androutsopoulos I, Paliouras G (2006) Spam filtering with Naive Bayes – Which Naive Bayes? *Proceedings of the third Conference on Email and Alnti-Spam (CEAS)*, pp. 27-28.

[33] Pakhomov S, Chacon D, Wicklund M, Gundel J (2011) Computerized assessment of syntactic complexity in Alzheimer's disease: A case study of Iris Murdoch's writing. *Behav Res Methods* **43**, 136-144.

[34] Sebastiani F (2002) Machine learning in automated text categorization. *ACM computing surveys (CSUR)* **34**, 1-47.

[35] Stamatatos E (2009) A survey of modern authorship attribution methods. *J Am Soc Inf Sci Technol* **60**, 538-556.

[36] John GH, Kohavi R, Pfleger K (1994) Irrelevant features and the subset selection problem. *Proceedings of the 11th International Conference on Machine Learning*, 121-129.