
Article:
Mirheidari, B., Blackburn, D., Harkness, K. et al. (4 more authors) (2017) Toward the Automation of Diagnostic Conversation Analysis in Patients with Memory Complaints. Journal of Alzheimer's Disease. ISSN 1387-2877

https://doi.org/10.3233/JAD-160507

Towards the automation of diagnostic conversation analysis in patients with memory complaints

Bahman Mirheidari[a], Daniel Blackburn[b], Kirsty Harkness[c], Traci Walker[d], Annalena Venneri[b,e], Markus Reuber[f], and Heidi Christensen[a]

[a] *Department of Computer Science, University of Sheffield, Sheffield, UK*
[b] *Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK*
[c] *Department of Neurology, Royal Hallamshire Hospital, Sheffield, UK*
[d] *Department of Human Communication Sciences, University of Sheffield, Sheffield, UK*
[e] *IRCCS Fondazione Ospedale San Camillo, Venice, Italy*
[f] *Academic Neurology Unit, University of Sheffield, Royal Hallamshire Hospital, Sheffield, UK*

Abstract

Background: The early diagnosis of dementia is of great clinical and social importance. A recent study using the qualitative methodology of conversation analysis (CA) demonstrated that language and communication problems are evident during interactions between patients and neurologists, and that interactional observations can be used to differentiate between cognitive difficulties due to neurodegenerative disorders (ND) or functional memory disorders (FMD).

Objective: This study explores whether the differential diagnostic analysis of doctor-patient interactions in a memory clinic can be automated.

Methods: Verbatim transcripts of conversations between neurologists and patients initially presenting with memory problems to a specialist clinic were produced manually (15 with FMD, and 15 with ND). A range of automatically detectable features focussing on acoustic, lexical, semantic and visual information contained in the transcripts were defined aiming to replicate the diagnostic qualitative observations. The features were used to train a set of five machine learning classifiers to distinguish between ND and FMD.

Results: The mean rate of correct classication between ND and FMD was 93% ranging from 97% by the Perceptron classifier to 90% by the Random Forest classifier.Using only the ten best features, the mean correct classification score increased to 95%.

Conclusion: This pilot study provides proof-of-principle that a machine learning approach to analysing transcripts of interactions between neurologists and patients describing memory problems can distinguish people with neurodegenerative dementia from people with FMD.

Keywords: Language, Dementia, Analysis, Machine Learning, Speech Recognition Software.

Introduction

The increasing number of people with dementia is one of the major concerns of health services [1, 2]. Memory clinics in the UK are experiencing a rising number of referrals [3], increasing the pressure on diagnostic services that were already struggling to provide timely assessments [4, 5]. The early differentiation of dementia from memory concerns unlikely to progress and not associated with a neurodegenerative disorder is highly desirable. It is, however, dicult to make this distinction clinically with sufficient accuracy. Existing cognitive screening tools for dementia lack sensitivity and specicity [3].

Memory complaints are common in the population across all ages. When memory concerns are sufficiently severe and persistent, and when no evidence of disease can be found, they are recognized as a functional symptom (i.e. a distressing somatic symptom associated with abnormal thoughts, feelings and behaviours). There are other functional neurological problems including "non-epileptic attack disorder", functional weakness or tremor, as well as functional non-neurological disorders, for instance characterised by abdominal or chest pain ("irritable bowel syndrome" and "non-cardiac chest pain") [6, 7]. Functional Memory Disorder (FMD) has been defined as a disorder in which people present with (potentially) reversible memory complaints, and which is thought to be caused by emotional or psychological factors [8, 9].

Language production is altered in the early stages of

neurodegenerative disorders (ND), and affects many aspects of language including object naming, noun production and rates of verb usage [10, 11]. Based on these observations and on previous research in which interactional observations were used as a differential diagnostic tool in patients presenting with seizures [12, 13], a recent study employed the qualitative methodology of conversation analysis (CA) to describe aspects of patient communication behaviour during first encounters with doctors in a memory clinic, which could help with the distinction of neurodegenerative memory disorders (ND) and (non-progressive) FMD [3, 14].

Traditional or 'manual CA' requires a number of steps including audio/video recording, detailed transcription and completion of a qualitative analysis by a trained expert. This process is expensive and time consuming and cannot be scaled up easily for routine clinical use. One way forward could be 'automatic CA' where dedicated speech technology software is used to analyse recorded interactions. Automatic CA is an emerging and challenging area of research that involves a number of disciplines [15]. It aims to automate the steps outlined above using automated speech recognition (ASR), speaker diarization (who is speaking when), and spoken language understanding (SLU). As the patient-doctor interactions are natural and largely unstructured, there are additional complexities including how to handle turn-taking, overlapping speech, prosody, sentence boundaries, as well as how to cope with dysfluencies or hesitations, and other paralinguistic data such as emotional content [16, 15].

To date, the automatic analysis of "natural" interaction (i.e., conversational speech) has not been used to support the differential diagnosis of memory problems. However, work has been carried out using machine learning techniques to identify signs of dementia in patients' speech and language. For instance, researchers have attempted to extract different types of features (e.g., acoustic, semantic and lexical) from speech samples of people with dementia produced in response to specific experimental prompts [17, 18, 19].

Several studies have shown that relatively high correct classification rates can be achieved when such methods are used to distinguish between individuals with dementia and healthy controls. In contrast, the diagnostic performance drops when attempts are made to classify between different types of dementia [20, 21] or between groups with more symptom overlap (such as patients with mild cognitive impairment (MCI) and patients with dementia [22]).

Recent research [23, 24] has used speech samples from the Dementia Bank corpus (capturing patients with AD, vascular dementia, MCI and healthy controls) to predict changes in patients' Mini Mental State Examination (MMSE) scores over time. The researchers automatically extracted a wide range of features from manually produced transcripts of the audio files. The prediction had a high level of accuracy, but it is unclear whether their results could be replicated by a fully automated ASR system. What is more, the distinction between AD and healthy controls represents much less of a diagnostic challenge in clinical practice than the differentiation of those with neurodegenerative dementias, MCI and age-matched adults with non-progressive memory complaints.

Materials and Methods

Participant recruitment and assessment

All participants recruited for this study were newly referred between October 2012 and October 2014 to the neurology-led memory clinic at the Royal Hallamshire Hospital in Sheffield, United Kingdom, with concerns about their cognition. Potential participants were sent information about this study prior to their appointment in the memory clinic. They were routinely encouraged to bring someone along to their memory clinic appointment if possible (accompanying person, AP). Written informed consent was obtained from all patients and accompanying persons by a member of the study team prior to their encounter with a neurologist. Participants and APs were only given the opportunity to consent if they had capacity to make their own decision about participation in the study and used English as their first language. Participants, whose diagnosis remained uncertain and those whose cognitive problems were considered to be due to other causes than ND or FMD, were excluded.

Participants were investigated and followed up by Consultant Neurologists specialising in memory disorders according to clinical need. All participants underwent MRI brain imaging and cognitive screening using the Addenbrooke's Cognitive Examination Revised (ACE-R). Participants underwent detailed neuropsychological testing with a neuropsychological battery which included the Mini Mental State Examination [25], tests of short and long term memory (verbal and non-verbal) [26], tests of abstract reasoning [27, 28], tests of attention and executive

function [29], language comprehension, naming by confrontation, category and letter fluency [30].

The diagnosis of Neurodegenerative Disorder (ND) was made according to standard criteria. Participants were attending their first ever appointment in the memory clinic and were mostly in the early disease stages. A few patients, however, were already at the moderate AD severity stage at the time of their first presentation. Alzheimer's disease was diagnosed according to the NINCDS-ADRDA criteria [31]. Patients' diagnoses were reached by a multidisciplinary team's consensus which took into account clinical history, neurological examination, neuropsychological scores and neuro-radiological findings. A diagnosis of mixed dementia (AD plus vascular cognitive impairment was made if moderate to severe small vessel ischaemic changes or cortical infarctions were present on MRI brain imaging). The diagnosis of behavioural variant Frontotemporal Dementia (bvFTD) was made according to Rascovsky criteria [32]. A diagnosis of amnestic MCI was made according to the criteria proposed by Petersen *et al.* [33]. We did not use biomarkers for amyloid or neurodegeneration (tau or FDG PET) because these tests are currently not available at our institution for routine assessment and because they are not widely used for clinical decision making in the NHS.

The diagnosis of FMD was based on the criteria proposed by Schmidtke *et al.* with the exception of the age cut-off of <70 years (subjective memory complaints for >6 months excluding cognitive deficits in the context of major psychiatric illness, absence of neurodegenerative disorder such as dementia or Mild Cognitive Impairment-MCI) [8]. We considered the age criterion overly restrictive because there have been many previous reports of cases of 'functional' (non-progressive) memory problems in people aged over 70. We excluded patients with long lasting or active depression Patient Health Questionnaire-9 (PHQ9) score of > 15. Participants were screened for Generalised Anxiety Disorder using the GAD7 but not excluded from the study on the basis of GAD7 scores (the exclusion of patients with depression but the inclusion of those with anxiety disorders is in line with the diagnostic criteria for FMD proposed by Schmidtke *et al.* [8]).

Memory Clinic Assessment

The participating doctors were encouraged to adhere to a communication guide, which had been developed in close cooperation with these clinicians and was based on their routine practice. Neurologists were guided to start their history-taking with an open enquiry, not explicitly directing patients to talk about their memory problems. They were encouraged to maximise patients' opportunities to produce an account of their own concerns and to minimise interruptions. After this open beginning, neurologists were asked to prompt further extended talk from patients by encouraging them to give an example of when their memory let them down. Finally, the communication guide listed some specific enquiries (such as who was more concerned about the memory difficulties, the patient or others). The ACE-R was carried out after the history-taking and not recorded or analysed.

Automatic conversation analysis system

A fully automatic CA system would compose of several units including the ASR, the speaker diarization, the feature extraction and the machine learning classifier (Figure 1). For the purposes of this pilot study, it was presumed that the transcripts of conversations produced by an ASR and diarization tool are close-to-perfect (the manual transcripts). The present study focuses exclusively on the CA-style feature extraction and the machine learning classifier units of the automatic CA system.

First, an audio file containing a recording of the conversation is entered into an automatic speech recognition (ASR) system. ASR uses computer science, machine learning, linguistic and signal processing techniques to generate a string of words from the audio recordings of human speech. The audio can be either stored in files or captured directly from microphone/telephone/mobile phone signals. The output of the ASR does not normally include information about the speaker identity nor any annotation of which words are spoken by whom. Speaker diarization tools are developed to provide this additional information. Diarization techniques first identify the speech and non-speech (silence, music, background noise, etc.) portions of the input audio stream, then, processing the speech parts of the input streams, they identify the speaker of each segment. The output of the ASR and diarization module consists of the text relating to the input utterance, the speaker of the utterance, and start and end time of the speech.

The output of the ASR+Diarization is passed on to the CA-style feature extraction unit which extracts certain features from the output. For instance, using the start time and end time of each turn of the conversation, the average length of the turn for a

specific speaker can be calculated. Some features may require further techniques such as text processing, natural language processing (NLP) and spoken language understanding (SLU). Extracting the number of unique words of a patient in a conversation is an example of using text processing, while the answer to a particular question in a conversation, is an instance of a feature extracted by SLU. SLU techniques focus on concepts rather than words, e.g. "I don't know", "I have no idea" and "don't know" all belong to the same conceptual category or meaning, although they are linguistically different.

Table 1 indicates which parts of the automatic CA system would be required to identify and analyse the qualitative features described in previous studies.

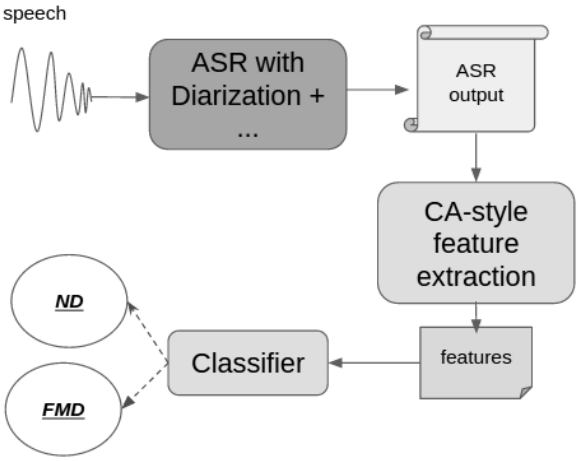

Figure 1. Automatic conversation analysis system: 1.Speech is processed by automatic speech recognition (ASR) with diarization 2.The output of the ASR and diarization tool is subjected to a CA-style feature extraction module 3. Finally, a machine learning classifier uses the identified features to classify whether the conversation belongs to the ND or the FMD category.

Pre-processing of transcripts

Transcripts were processed further by conversion to XML (Extensible Markup Language) files containing only raw text, grammatical punctuations, pauses, and some non-verbal information such as 'patient turns to other' (part c in Figure 2). XML is a standard markup language in a format which is easily readable by both humans and machines [34]. The XML file keeps information inside tags which may contain other inner tag(s) and/or attributes (with values). For instance, the whole conversation is divided into consecutive tags of 'turn' where each turn includes the attribute's start time, end time (e.g., the turn starts at time 741.30 second and ends at time 748.10) and speaker id (PAT, NEU, APS - patient, neurologist and accompanying person respectively), Figure (2c). Phrases are inner tags for the turns i.e., each turn can be split into one or more 'phrase' tags. Note that in the table in Figure (2a), timing information has been omitted for the sake of simplicity and numbers have been used instead (turn 1, 2, etc.). In the original transcripts, the start time of each turn was provided, but not the end time. Therefore, the end time was considered equal to the start time of the next turn.

Extracted Features

In the process of translating qualitative features from the earlier CA study [3, 14], it became apparent that, in most cases, several complementary programmable features had to be combined to generate a reasonably close translation of a qualitative observation. Table 2 shows how the six key features described qualitatively were translated into 17 features suitable for automatic CA. In addition, we defined five potentially diagnostic features suited for automatic CA which focussed on the interactional contributions of the neurologist, but which were not based on any previous qualitative findings (see Table 2).

To calculate some of those features, a common natural language (NLP) approach known as the Bag-of-Words (BoW) model [35] was used. This technique underpins many search engines (like Google) and is supported by numerous NLP packages (e.g. NLTK [36]). The BoW ignores the order of words, punctuation, commonly used words in English (such as 'the', 'a', 'this' ; since they are not specific they cannot help with searches), and trims verbs to their stems. For instance, for the clause 'He wanted to get a new job', the BoW would contain the words: 'want', 'new', and 'job'.

The detection of most of the features depends on an automatic way of identifying turns, i.e., splitting the conversation into questions and answers. This is a relatively hard task. However, this study is based on the automated analysis of a small number of highly structured conversations (30 conversations in total), in which new topics are almost exclusively initiated by the clinician. This means we were able to use a far simpler topic detection method relying on the detection of particular words or phrases in a turn. This facilitates the extraction of features aiding the identification of the role of the APs (F1). Features such as the number of turns, the average length of turn and the average number of unique words produced by the patient and the AP can be used individually

Table 1. Summary of the six key linguistic features extracted by Elsey *et al.* [3] with an additional computer engineering view on how to extract the features automatically utilizing current speech and NLP technologies.

Linguistic feature (Elsey *et al.*)	**Findings in FMD group**	**Finding in ND group**	**Computer engineering view (current study)**
F1.Accompanying persons (AP)	Providing confirmation when prompted by patient	Main spokesperson for the patient	Needs ASR + diarization + text processing tool
F2.Responding to "who's most concerned"	Typically the patients themselves	Typically others or "I don't know" answer	Needs ASR + diarization + SLU
F3.Patient recall of recent memory failure	Patient able to provide detailed account	Patient has difficulties with answering, "don't know" or gives general answers ("all the time")	Needs ASR + diarization + SLU
F4.Inability to answer	Infrequent use of "don't know" answers	Frequent use of "I don't know" or nonverbal behaviours like head turning, encouraging AP to answer	Needs ASR (with calculating the silence length) + diarization + SLU + motion detection camera
F5.Responding to compound questions	Able to answer all parts of the question	Only answering a single part of the question	Needs ASR + diarization + SLU
F6.Patients' elaborations and length of turns	Frequent, unprompted elaborations, long turns	Very little unprompted elaboration, short turns	Needs ASR + diarization + NLP

to determine whether the patient or the AP talks more. A total of six features are defined to select the dominant speaker: the number of turns in the conversation (PatNoOfTurns and APsNoOfTurns), the average length of the turns (PatAVTurnLength and APsAVTurnLength), and the average unique number of words in the whole conversation (PatAVUniqueWords and APsAVUniqueWords).

To extract information related to who is the most concerned about the patient's condition (F2), the topic detection approach described above is used first to identify the question, and subsequently assess the associated answer to determine whether the patient has replied that they are the most concerned (in effect answering "me") or not (PatMeForWhoConcerns). Since not all the patients were asked this question, the feature actually had three possible values: "yes", "no", and "not available".

F3 relates to the question when patients last noticed a problem with their memory. Patients with ND were found to give three different types of answer to this question: providing mostly empty words, answering with a lot of hesitation or gaps in the speech, or answering something to the effect of 'all the time'. Therefore, three features were defined to capture the answer to this question: number of empty words in the response (PatFailureEx-ampleEmptyWords), the average length of silences within the utterances (PatAVPauses), pause for failure example (PatFailureExampleAVPauses), and replying 'all the time' (PatFailureExampleAllTime).

In order to extract the feature "inability to answer", five different features were defined. The feature PatDontKnowForExpectation indicates that either the patient has replied "I don't know" or used a similar phrase in response to the question about what expectations they had when they came to the clinic. Elsey *et al.* also described "don't know" responses at other points of the interaction as diagnostically meaningful, although they differentiated between different types of this particular response: contextualised "don't knows" in which the speaker provides appropriate information addressing parts of a question but identifies particular aspects s/he is unable to answer, or non-contextualised "don't know" responses in which no attempt is made to provide a more detailed reply to any aspect of a question. To improve the diagnostic contribution of "don't know" statements, we therefore did not only count these utterances (PatAVNoOfDontKnow), we also coded additional information sometimes associated with these words (such as patient turns head to the AP encouraging them to answer the question instead of the patient). Similarly, we coded head shaking (translated into the feature PatAVNoOfShakesHead). Other important features, which may be helpful in determining the meaning of "don't know" statements, are the average number of filler words like "I mean", "I see"

a) Manual CA

056 (dementia, accompanied)		
1	Neu	How's er:reading, writing, spelling?
2	Pat	Erm(.) <reading >(.) I read an awful lot(.) however, I have-and the only
3		way I've noticed it is, well we've got a three year old grandson and I=
4	AP	=Oh yeah

b) Transcript file

```
(0:12:21.3) Neu: How's er reading, writing, spelling?
(0:12:28.1) Pat: Um, reading, I read an awful lot, however, I have, and the only
way I've noticed it is, well we've got a (laughs) three year old grandson and I.
Oth: Oh yeah.
```

c) XML file

```
<?xml version='1.0'?>
<conversations>
    ...
    <turn starttime="741.30" endtime="748.10"
        speaker="NEU056">
        <phrase type="verbal">How's er reading,
            writing, spelling?</phrase>
    </turn>
    <turn starttime="748.10" endtime="759.50"
        speaker="PAT056">
        <phrase type="verbal">Um, reading, I
            read an awful lot, however, I have,
            and the only way I've noticed it
            is, well we've got a</phrase>
        <phrase type="others" value="LAUGHS"/>
        <phrase type="verbal">three year old
            grandson and I.</phrase>
    </turn>
    <turn starttime="759.50" endtime="759.50"
        speaker= APS056">
        <phrase type="verbal">Oh yeah.</phrase>
    </turn>
    ...
</conversations>
```

Figure 2. a) Qualitative CA for patient 056 from Elsey *et al.* [3]. The prefixes Pat (patient), Neu (neurologist) and APs (accompanied person(s)) identify the speaker of the utterance at each turn. At turn 1, the neurologist asks the question "How's er: reading, writing, spelling?" and in turn 2, the patient answers "Erm(.) <reading >(.)", and so on. b) The corresponding transcript file with start time (e.g. 12 minutes and 21.3 second for the first turn), speaker (Neu, Pat, or APs) and text with additional non-verbal information (e.g. laughs in parentheses). Some of the CA symbols have been removed from the transcript. c) The converted XML file in which the conversation is split into tags of 'turn' and each turn contains attributes of start time, end time, speaker (with a suffix indicating the ID of the recording; here 056). Inside the turn tag, there are verbal, and other (non-verbal) phrases (eg. LAUGHS).

(PatAVFillers), the average number of empty words such as "er", "em" (PatAVEmptyWords) and the average number of words in a turn (ignoring very common words such as "a", "the", "that", PatAVAllWords)

In their responses to compound (multi-part) questions (F5), ND patients typically failed to answer all parts of the question so the neurologist had to repeat the question in the following turn. This is captured by feature AVNoOfRepeatedQuestions which takes into account parts of compound questions which were not answered by the patient straight away.

The lack of elaboration of answers by patients with ND was captured by the features PatNoOfTurns, PatAVTurnLength, and PatAVUniqueWords.

In recognition of the conversation analytic axiom of the co-construction of interaction by all speakers, we also extracted three features based on the contributions of neurologists. Although the differential diagnostic value of the neurologists' contribution has not been studied explicitly by Elsey *et al.*, it has been identified as a conversational observation of potential value by others [11, 37]. Similar to the APs and patient features, NeuNoOfTurns, NeuAVTurnLength, and NeuAVUniqueWords were identified. Finally, the feature AVNoOfTopicsChanged takes into account the average number of different topics discussed by the neurologist and patient throughout the conversation.

The extracted features can be divided into four different types: acoustic, lexical, semantic and visual

Table 2. Linguistic features from Table 1, and corresponding features extracted from the transcripts by automatic CA. Prefixes: Pat:patient, Neu:neurologist, and APs:accompanying person(s).

Linguistic feature (Elsey *et al.*)	**Corresponding extracted feature(s)**
F1.Accompanying persons (AP)	Number of turns (**1.APsNoOfTurns** , **2.PatNoOfTurns**; average length of turn ([sec]) (**3.APsAVTurnLength**, **4.PatAVTurnLength** ; average unique words in a turn (**5.APsAVUniqueWords**, **6.PatAVUniqueWords**)
F.Responding to "who's most concerned"	patient answered "me" (**7.PatMeForWhoConcerns**)
F3.Patient recall of recent memory failure	Number of empty words (**8.PatFailureExampleEmptyWords**); average length of pauses (**9.PatFailureExampleAVPauses**); used "all the time" (**10.PatFailureExampleAllTime**)
F4.Inability to answer	Patient replies "I don't know" to the question about their expectations of the memory clinic appointment (**11.PatDontKnowForExpectation**); frequency of "don't know" responses in combination with turning to AP (**12.PatAVNoOfDontKnow**); average instances of head shakes (**13.PatAVNoOfShakesHead**); average number of filler words (**14.PatAVFillers**); average number of empty words (**15.PatAVEmptyWords**); average number of common words (**16.PatAVAllWords**)
F5.Responding to compound questions	Average number of repeated questions (**17.AVNoOfRepeatedQuestions**)
F6.Patients' elaborations and length of turns	Patient's average unique words in a turn (**6.PatAVUniqueWords**, **4.PatAVTurnLength**)
Role of the neurologist (Not in Table 1)	Number of turns (**18.NeuNoOfTurns**); length of turns([sec]) (**19.NeuAVTurnLength**); average number of unique words (**20.NeuAVUniqueWords**); average number of topics discussed (**21.AVNoOfTopicsChanged**); average length of pauses by patient (**22.PatAVPauses**)

(non-verbal). Table 3 lists all features.

The feature type categorisation is based on the type of information used by the computer when detecting those features automatically. For instance, the information about the turns, such as the length of the turn, can be extracted by the diarization tools, which provides information about when the speaker talks, which in turn enables us to extract the start time and the length of turn. The diarization tool uses the acoustic signals of the input utterances, and the turn-related features are, therefore, categorised as 'acoustic'.

Table 3. Types of extracted features: acoustic, lexical, semantic and visual-conceptual.

Type	**Features**
Acoustic	APsNoOfTurns PatNoOfTurns NeuNoOfTurns APsAVTurnLength PatAVTurnLength NeuAVTurnLength PatAVPauses
Lexical	PatAVUniqueWords NeuAVUniqueWords APsAVUniqueWords PatAVAllWords
Semantic	PatMeForWhoConcerns PatFailureExampleEmptyWords PatFailureExampleAllTime PatDontKnowForExpectation PatAVFillers PatAVEmptyWords AVNoOfRepeatedQuestions AVNoOfTopicsChanged
Visual-conceptual	PatAVNoOfShakesHead PatAVNoOfDontKnow

Machine learning and classifiers

Machine learning aims to construct algorithms capable of learning to detect patterns in input data, allowing the system to generate a model (normally a statistical model) in order to make decisions or predictions of previously unseen (new) data [38, 39, 40]. Decision making or prediction is carried out in two forms: assigning a class or category to the new data (known as classification; e.g., in Figure 1, the input speech either is from a patient with FMD or ND, so those will be the target classes for the classifier), or allocating a value to new data (regression, e.g.,

predicting the temperature of London tomorrow at 10am based on the forecasting data gathered from previous days or years) [40].

There are normally two stages in machine learning: training and testing. In the training stage, using different input data (set of features) and the associated classes or values (target output class) for the data, the learning algorithms learn the relationship between input and output data. In the testing stage, using the learned patterns from the training stage, the constructed model is asked to generalise and assign classes or values to new, unseen input data. The algorithm or programme which performs the classification task is called a classifier.

There are several standard machine learning classifiers, however, choosing the best classifier for a given dataset is a challenging task, because each one has advantages and disadvantages, depending on factors such as the number of samples of training and testing data, and the variances of the different features in the data. Therefore, a very common methodology is to try several classifiers and use a validation approach to find the best classifier for a particular dataset.

The focus of this study was the differentiation between patients with ND and FMD, so a binary machine learning classifier was used. The 'Scikit-learn' [41] is a Python library with a wide range of machine learning classifiers. From this library, five standard machine learning classifiers were chosen: Support Vector Machine (SVM) with linear kernel, Random Forest, Adaptive Boost (AdaBoost), Perceptron, and Stochastic Gradient Descent (SGD).

Results

Participants

Of 353 patients who received the invitation, 148 were eligible. 36 declined to take part. 112 gave consent to take part in the study. Three withdrew their consent subsequently, leaving 109 who completed the study. 19 participants received a diagnosis of ND. 4 were later removed from further analysis because they lacked detailed neuropsychology results. 30 participants were diagnosed with FMD, 26 with depressive pseudodementia, 12 with vascular cognitive impairment not related to dementia and the diagnosis remained uncertain in 22 cases. This study is based on analyses of the 15 patients with ND and the first 15 patients with FMD whose conversational data were analysable and in whom the diagnoses could be established with sufficient certainty (See Figure 3 - CONSORT diagram). The 15 cases categorized as ND comprised eight with Alzheimer's Dementia (AD), three mixed AD and vascular, two amnestic MCI, two with Frontotemporal Dementia (FTD) (Demographic data are shown in Table 4).

The participants with ND were not significantly different in age, although there was a trend towards being older. Their mean scores on cognitive screening (Addenbrooke's Cognitive Examination - ACE) and detailed neuropsychological testing were below cut-off (see table 4) and all significantly lower than the FMD group. In two cases of FMD there was co-morbid sleep disturbance and one was later diagnosed with obstructive sleep apnea. One case had co-morbid fibromyalgia.

Fifteen participants with FMD; one participant with FMD had MRI brain scan reported as possible atrophy. This patient was followed up for 18 months. The Montreal Cognitive Assessment (MoCA) was 26/30 at follow-up (prior ACE 85 and MoCA 22). They were functioning normally and holding a busy job.

Fifteen cases of ND; all had formal neuropsychology. Two out of fifteen cases had normal structural scans but one had abnormal Single-Photon Emission Computed Tomography (SPECT). The other was seen for follow-up at 12 months, ACE-R had decreased from 87 to 82, and findings were clinically consistent with AD.

Validation

In this study, we used a common validation technique, recommended for a classification task with a small number of data samples, the 'leave-one-out' approach. In this approach, in a loop over n (the number of total samples), each time one of the samples is taken out and used for testing, while the rest of the data $(n-1)$, is used for training. The average score over all the tests determines the accuracy of the classifier. Table 5 displays the overall accuracy in percentage for the five selected classifiers using all 22 features extracted from the transcripts. The best score was achieved by the Perceptron classifier with 97% accuracy, while the minimum score was achieved by the Random Forest classifier with 90% accuracy. The mean correct classification score of all classifiers was 93.2% (standard deviation is 2.5%).

Feature selection

Generally, the best features to use in automated classification approaches are complementary and highly

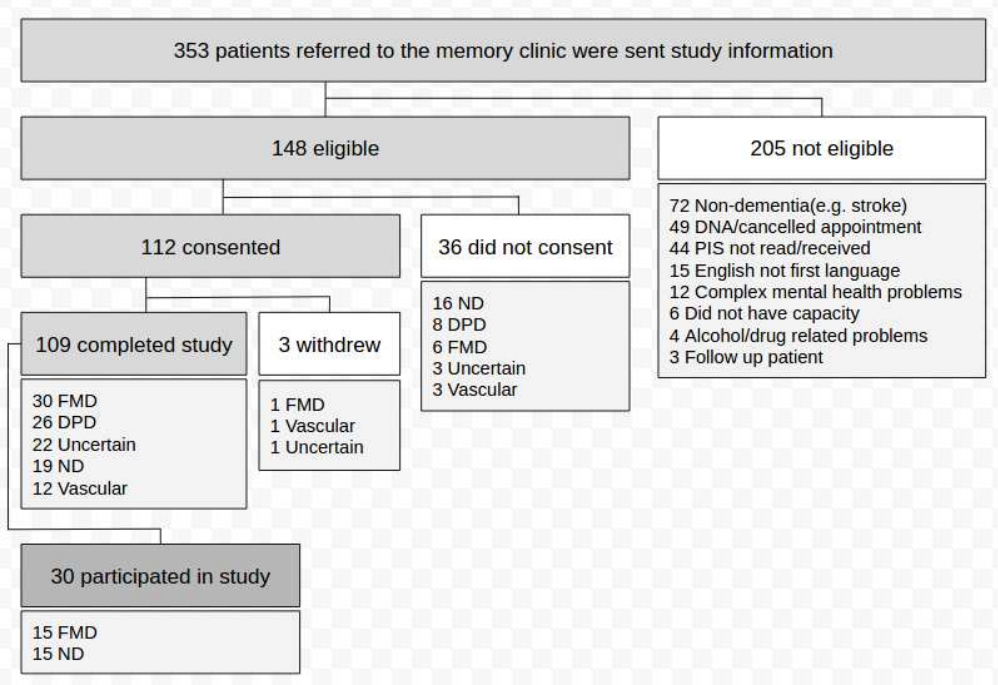

Figure 3. CONSORT flow diagram.

discriminative for the task at hand. In practice though, it is common for two types of features to exhibit a high degree of interdependence, and a process of feature selection is often beneficial. This makes the machine learning model simpler (fewer features need to be extracted), regulates the variance amongst the extracted features and, more importantly, reduces the risk of overfitting (many features do not necessarily yield better classification, but rather make the final prediction too dependent on a specific dataset [42]). One approach is to consider the input data (disregarding the output classes) with the aim of identifying those features with the greatest variance and diversities using statistical tests such as t-tests. Another approach considers the outputs of the classification task in order to find the most discriminative features. Feature selection in this way depends on the amount a particular feature contributes to the classification. Some classifiers, such as those that are based on trees, automatically use feature contribution for the classification task, therefore they have a built-in ranking which can show the importance of features. For linear classifiers, Recursive Feature Elimination (RFE) (see [41] for more details) is a common approach to selecting the top features. RFE finds the most important features by examining how eliminating each feature from the feature set affects the classification accuracy. One by one, the feature making the smallest contribution is eliminated and the accuracy of the remaining features is evaluated. Elimination continues until all features have been eliminated. Reverse order elimination shows the importance of the features in the classification task.

For the tree-based classifiers (AdaBoost and Random Forest) the built-in ranking was employed and for the other linear classifiers, the RFE technique was used to identify the best features. The top 10 features overall were selected by combining the feature rankings of five classifiers. Table 6 lists the most important features contributing to the classification. The top five features were the average number of unique words used by the neurologist, accompanied person's number of turns, the average number of unique words used by the patient, the average turn length for the patient, and the average number of repeated questions.

There are other approaches such as component analysis e.g., the PCA (Principle Component Analysis) which can be used to reduce the dimensionality of the features. However PCA is better suited to reducing very large datasets (e.g., with hundreds of features) and also, by using feature selection methods directly affected by the classifier in question, we ensure we identify the most important features for the task at hand. This also enables us to arrive at a subset of features that needs extracting as opposed

Table 4. Demographic information of the participants.

	FMD (n=15)	**ND (n=15)**	**Mean**	**Cut Off**	**Max Score**	**P value**
Age	$57.8(+/-2.02)$	$63.73(+/-2.29)$	N/A	N/A	N/A	$p = 0.06$
Female	60%	53%				ns^*
ACE-R	$93.0(+/-1.4)$	$58.27(+/-5.21)$		88	100	$p < 0.0001$
MMSE	$28.87(+/-0.19)$	$18.79(+/-1.97)$	28.88(1.28)	26.32	30	$p < 0.0001$
PHQ9	$5.6(+/-1.02)$	$5.25(+/-2.04)$		5	27	ns
GAD7	$4.73(+/-1.23)$	$4.75(+/-1.52)$		5	21	ns
CF	$19.8(+/-0.11)$	$17.15(+/-0.93)$	19.65(0.63)	18.39	20	$p = 0.0052$
VPA	$16.87(+/-0.74)$	$5.85(+/-0.94)$	14.81(3.76)	7.29	24	$p < 0.0001$
P&PT	$51.13(+/-0.19)$	$44.50(+/-2.49)$	51.23(0.82)	49.59	52	$p = 0.0063$
Rey's CF	$34.0(+/-0.44)$	$21.42(+/-3.02)$	33.70(2.30)	29.1	36	$p < 0.0001$
SF	$52.73(+/-2.91)$	$23.77(+/-4.03)$	59.81(13.17)	33.47	N/A^{**}	$p < 0.0001$
PF	$41.2(+/-3.02)$	$19.15(+/-3.69)$	45.58(12.05)	21.48	N/A^{**}	$p < 0.0001$
DS	$6.73(+/-0.33)$	$4.54(+/-0.48)$	6.76(1.48)	3.8	9	$p = 0.0007$
VCA	$13.2(+/-0.2)$	$10.08(+/-0.97)$	13.77(0.51)	12.75	14	$p = 0.0023$
TT	$34.97(+/-0.27)$	$26.50(+/-1.89)$	34.67	1.03	36	$p < 0.0001$
PM	$15.07(+/-0.92)$	$5.25(+/-1.1)$	12.37	2.08	25	$p < 0.0001$

Legends: ACE-R: Addenbrooke's Cognitive Examination - Revised; MMSE: Mini Mental State Examination; PHQ9: Patient Health Questionnaire-9; GAD-7: Generalised Anxiety Assessment 7; CF: Confrontational Naming; VPA: Verbal Paired Associates; P&PT-Pyramid & Palm Trees; Rey's CF: Rey's Complex Figure; SF: Semantic Fluency; PF: Phonemic Fluency; DS: Digit Span; VCA: Visuoconstructive Apraxia; TT: Token task; PM: Prose Memory. trials.
*: not significant
**:For the CF and PF tests there is no maximum score as it depends on individuals' word production speed within the time limit of three one minute trials. We have included all the maximum scores on the cognitive tests apart from GAD7 and PHQ9 where we have included the minimum to reflect a score if no depression or anxiety were present.

Table 5. Classifiers' scores using all 22 extracted features.

Classifier	**Score** (%)
Linear SVM	93
Random Forest	90
AdaBoost	93
Perceptron	97
Linear via SGD	93
AVG (STD)	*93.2(2.5)*

to PCA-based reduction which would still require us to extract all features prior to dimensionality reduction.

Using only the top 10 features instead of all 22 features resulted in a better performance for most of the five nominated classifiers. The mean accuracy of correct diagnosis prediction improved to 95.4% across all classifiers with standard deviation of 2.2%. While the correct classification rate of the Perceptrion dropped from 97% to 93%, the accuracy rate for the Linear SVM, AdaBoost and the linear via SGD rose from 93% to 97% (see Figure 4).

Table 6. Top 10 features with the highest contributions for the classification between the ND and the FMD patients.

Rank	**Feature Name**
1	NeuAVUniqueWords
2	APsNoOfTurns
3	PatAVUniqueWords
4	PatAVTurnLength
5	AVNoOfRepeatedQuestions
6	PatFailureExampleEmptyWords
7	PatAVFillers
8	PatAVAllWords
9	PatMeForWhoConcerns
10	PatAVPauses

Feature type importance

In order to identify the relative diagnostic contribution of individual feature types, the classification task was repeated using only acoustic, only lexical, only semantic and only visual-conceptual features. The results are presented in Figure 5; however, the importance of feature type depends on the classifier itself to some degree. In brief, lexical features are the least important with a classification score rang-

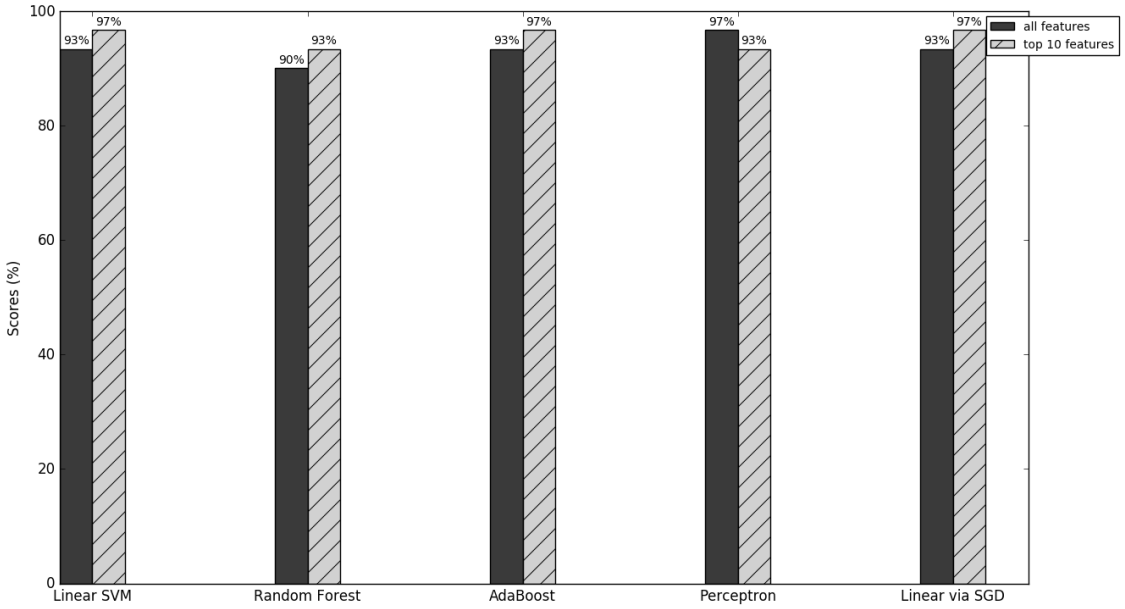

Figure 4. Comparison of classification rates using individual classifiers based on all features and the top 10 features.

ing from 73% for the Random Forest to 53% for the Perceptron, whereas the visual-conceptual features are the most important feature types with over 90% contributions in classification for the Random Forest and 83% for the linear SVM and AdaBoost classifiers. Semantic features are the second most important feature types and the acoustic features are the third.

Discussion

The early diagnosis of neurodegenerative disorders and their distinction from normal ageing or FMD is a challenging clinical task. The Royal College of Psychiatrist's audit of memory clinics revealed a fourfold increase in referrals from primary care to specialist memory clinics between 2011 and 2013, and a 31% increase in referrals between 2013 and 2014 [43]. This rise in the referral rate has been associated with an increase in the proportion of patients referred with subjective memory concerns but no evidence of dementia [4, 5, 44]. Currently the decision to refer is based on a GP's interpretation of the history given by patient and informant (such as partner, friend or family member) and the result of short screening tests. Although these tests have a high sensitivity, they have a low specificity for dementia [45, 46]. Our study suggests that automated conversation analysis has the potential to improve the screening and triage procedures for patients with possible ND. The improvement of case selection for referral to specialist clinics would mean that those at high risk of developing dementia could be seen more quickly, whilst those with FMD could be reassured at an earlier stage in the clinical management pathway. Although further work is required to develop our method into a screening tool that could be deployed in primary care, the approach described here has the advantage of being non-invasive and usable in a wide range of healthcare settings.

Previous studies have already demonstrated that the qualitative examination of interaction between doctors and patients using CA can help with the differentiation of memory problems due to ND or FMD [3, 14]. However, formal CA is based on an expert post-hoc assessment of video/audio recordings and transcripts intended to capture interaction in great detail. This study explored whether the insights gained by expert qualitative study of detailed transcripts can be used to develop an automated screening process for ND. We tested a simplified version of automatic CA, focussing on machine learning and classification. Five standard machine learning classifiers were used. The best classifier was the Perceptron, which was very accurate (up to 97%), while

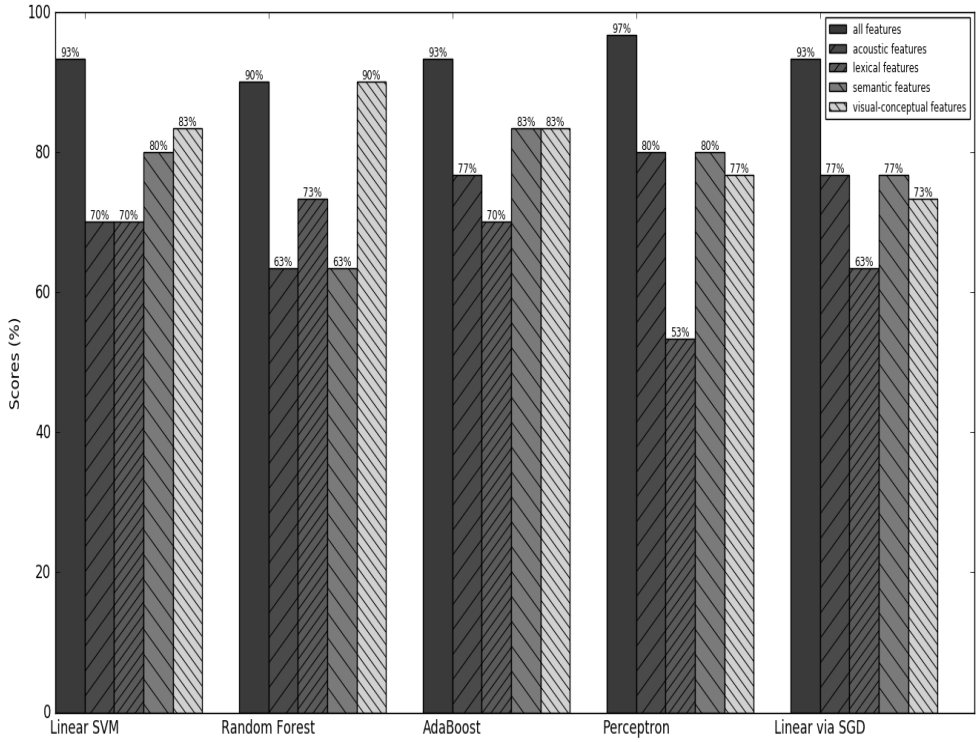

Figure 5. Classifiers' scores for different types of features acoustic, lexical, semantic and visual-conceptual, as well as all features.

the mean accuracy for the five selected classifiers was 93%. Use of the top 10 features resulted in improved overall performance than use of all initially defined features. Visual-conceptual and semantic features were most important for the classification, suggesting that an optimised screening process does not only need to take account of verbal information but also of visual features.

In addition to demonstrating the potential for automation of the diagnostic analysis of interactional data, this study provides independent validation of the qualitative, microanalytic sociological methodology of CA which provided the basis for the definition of the automatically extractable features used in the present study. In addition, the present study has identified some new interactional features with diagnostic potential now requiring further in-depth analysis using methods such as CA: quantitative features extracted from the contributions which APs and neurologists made to the conversation were amongst the top ten diagnostic features. The contributions of these individuals were not studied in the previous qualitative studies of memory clinic encounters by Jones *et al.* [14] and Elsey *et al.* [3]. Whilst the interactional role of APs in clinic conversations requires more research, the conversational role of caregivers to people with dementia (i.e. individuals with more significant cognitive problems than those exhibited by the patient group described here) has been studied by Perkins *et al.* [37], focussing on turn taking, repair and topic management. They found that caregivers had a key role in successful conversations. For instance, caregivers used touch, gaze and the patient's name before talking, to achieve better responses from patients. Greater familiarity between patient and caregiver reduced dysfluencies, mishearing and misunderstanding, while unfamiliarity between the interviewer and the patient resulted in fewer topic initiations.

It is possible that the differences in neurologists' communication behaviour in encounters with ND patients on the one hand and FMD patients on the other, which we picked up by automated CA in this study, are due to the fact that they became aware

of the diagnosis relatively early in the consultation. Future studies will need to examine whether less expert clinicians (for instance those working in primary care) would change their communication in similar ways and whether they could be made more aware of that fact that they are adjusting their conversational style (which could help with the diagnostic process).

This study has several limitations. Although the recruitment of patients first referred to a memory clinic with cognitive concerns increases the clinical validity of our findings, our recruitment method means that the findings cannot be readily generalised to patients complaining of memory problems in other settings, for instance in primary care. Furthermore, we were only able to analyse a relatively small number of conversations. The patients whose interactional behaviour we studied, however, represented two neurologically well-characterised groups. Importantly, our study did not compare patients with ND with healthy controls but with patients with FMD, enhancing the practical relevance of our findings. We assumed perfect accuracy of transcription by ASR, which is not an uncommon first step in this research area. Looking ahead, this part of an automated CA system will be one of the most difficult aspects and will need to be the focus of further studies. It is possible that features not described here would perform better diagnostically if less perfect transcripts than used in this study were employed in a fully automated diagnostic procedure. Furthermore, in this initial proof-of-concept study we focused on a relatively small number of features described by Elsey *et al.* [3]. There are, however, potentially many other distinctive semantic, acoustic and lexical features that could be extracted from audio or video recordings which may further improve the classification accuracy.

We acknowledge that definitive diagnosis in patients with memory disorders requires long term follow-up and post-mortem studies. We were also unable to use biomarkers in blood or cerebrospinal fluid, or amyloid PET to confirm clinical diagnosis. However, our studies were based on clinical assessments by Consultant Neurologists specialising in memory disorders, detailed neuropsychological examination in all ND cases and MRI brain imaging. We excluded all patients in whom any of the participating experts had any doubt about the categorisation in the ND or FMD groups.

Although we have used clearly defined diagnostic criteria we also acknowledge the difficulties which arise when our findings are compared to those of studies using different diagnostic labels or categories. For instance, there is overlap between labels of FMD and Subjective Cognitive Impairment (SCI) [47]. It is important to note the differences between the SCI and FMD concepts. SCI has been defined and used to identify a population at greater risk of developing AD and based on older adults (typically over 70 years of age) [48, 49]. In younger patients SCI is more closely associated with psychological distress than neurodegeneration [50]. In our study, we aimed to distinguish between a group of patients with memory problems thought not to be at increased risk of development of a neurodegenerative disorder and those in the early stages of dementia or very likely to develop dementia in the near future. The study by Schmidtke *et al.* [8] followed up 47 participants with FMD for an average of 20 months and found that symptoms persisted in 39 participants, resolved in 6 , and that one person had developed early stage AD. This study demonstrates that automatic linguistic and speech analysis can differentiate between this patient group and patients with ND with a high level of accuracy.

Despite its limitations, this study demonstrates the feasibility of translating interactional findings derived from the qualitative study of transcripts into features which can be automatically extracted and analysed. Our findings show that such an automated process has the potential to improve the early identification of patients at high risk of developing dementia. At the same time our study provides further support for the validity of CA, the qualitative method used to identify the diagnostic features our automated extraction and analysis method was trained to detect.

References

[1] Alzheimer's society :dementia uk update, https://www.alzheimers.org.uk/dementiauk, Accessed on October 22, 2015.

[2] Dementia research funding to more than double to 66m by 2015, http://www.theguardian.com/society/2012/mar/26/dementia-research-funding-to-double, Accessed on October 22, 2015.

[3] Elsey C, Drew P, Jones D, Blackburn D, Wakefield S, Harkness K, Venneri A, Reuber M (2015) Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics,

Patient Education and Counseling **98** 1071–1077.

[4] Bell S, Harkness K, Dickson JM, Blackburn D (2015) A diagnosis for 55: what is the cost of government initiatives in dementia case finding., *Age and Ageing* **44** 344–345.

[5] Larner AJ (2014) Impact of the National Dementia Strategy in a neurology-led memory clinic: 5-year data., *Clinical Medicine* **14** 216.

[6] Carson AJ, Brown R, David AS, Duncan R, Edwards MJ, Goldstein LH, Grunewald R, Howlett S, Kanaan R, Mellers J, et al. (2012) Functional (conversion) neurological symptoms: research since the millennium, *Journal of Neurology, Neurosurgery & Psychiatry* **83**, 8 842–850.

[7] Stone J, Carson A, Duncan R, Coleman R, Roberts R, Warlow C, Hibberd C, Murray G, Cull R, Pelosi A, et al. (2009) Symptoms unexplained by organic diseasein 1144 new neurology out-patients: how often does the diagnosis change at follow-up?, *Brain* awp220.

[8] Schmidtke K, Pohlmann S, Metternich B (2008) The syndrome of functional memory disorder: definition, etiology, and natural course, *The American Journal of Geriatric Psychiatry* **16**, 12 981–988.

[9] Metternich B, Schmidtke K, Hüll M (2009) How are memory complaints in functional memory disorder related to measures of affect, metamemory and cognition?, *Journal of Psychosomatic Research* **66**, 5 435–444.

[10] Bayles KA, Kaszniak AW (1987) *Communication and cognition in normal aging and dementia*, Taylor & Francis Ltd London.

[11] Hamilton HE (1994) *Conversations with an Alzheimer's patient: An interactional sociolinguistic study*, Cambridge, England: Cambridge University Press.

[12] Schwabe M, Reuber M, Schondienst M, Gulich E (2008) Listening to people with seizures: how can linguistic analysis help in the differential diagnosis of seizure disorders?, *Communication & medicine* **5**, 1 59.

[13] Reuber M, Monzoni C, Sharrack B, Plug L (2009) Using conversation analysis to distinguish between epilepsy and non-epileptic seizures: a prospective blinded multirater study, *Epilepsy Behav* **16**, 1 139–44.

[14] Jones D, Drew P, Elsey C, Blackburn D, Wakefield S, Harkness K, Reuber M (2015) Conversational assessment in memory clinic encounters: interactional profiling for differentiating dementia from functional memory disorders, *Aging & Mental Health* **7863** 1–10.

[15] Moore RJ (2015) Automated Transcription and Conversation Analysis, *Research on Language and Social Interaction* **48**, 3 253–270.

[16] Shriberg E (2005) Spontaneous speech: How people really talk and why engineers should care, *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech* 1781–1784.

[17] López-de Ipiña K, Alonso JB, Travieso CM, Solé-Casals J, Egiraun H, Faundez-Zanuy M, Ezeiza A, Barroso N, Ecay-Torres M, Martinez-Lage P, Martinez de Lizardui U (2013) On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer's disease diagnosis., *Sensors* **13** 6730–45.

[18] López-de Ipiña K, Solé-Casals J, Eguiraun H, Alonso J, Travieso C, Ezeiza A, Barroso N, Ecay-Torres M, Martinez-Lage P, Beitia B (2015) Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach, *Computer Speech & Language* **30** 43–60.

[19] Tóth L, Gosztolya G, Vincze V, Hoffmann I, Szatlóczki G, Biró E, Zsura F, Pákáski M, Kálmán J (2015) Automatic detection of mild cognitive impairment from spontaneous speech using ASR, *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech* .

[20] Jarrold W, Peintner B, Wilkins D, Vergryi D, Richey C, Gorno-Tempini ML, Ogar J (2014) Aided diagnosis of dementia type through computer-based analysis of spontaneous speech, *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* 27–37.

[21] Satt A, Sorin A, Toledo-Ronen O, Barkan O, Kompatsiaris I, Kokonozi A, Tsolaki M (2013) Evaluation of speech-based protocol for detection of early-stage dementia, *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech* 1692–1696.

[22] Thomas C, Keselj V, Cercone, Rockwood K, Asp E (2005) Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech, *Proceedings of the IEEE International Conference on Mechatronics & Automation* 1569–1574.

[23] Fraser KC, Meltzer JA, Rudzicz F (2015) Linguistic Features Identify Alzheimer's Disease in Narrative Speech, *Journal of Alzheimer's Disease* **49** 407–22.

[24] Yancheva M, Fraser K, Rudzicz F (2015) Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias, *6th Workshop on Speech and Language Processing for Assistive Technologies* .

[25] 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**, 3 198–98.

[26] Wechsler D (1997) *Wechsler Adult Intelligence Scale*, The Psychological Corporation., 3rd edition.

[27] Rey A (1964) *Lexamen clinique en psychologie*, Presses universitaires de France., 2nd edition.

[28] Raven JC (1995) *Coloured Progressive Matrices Sets A, Ab, B. Manual Sections 1 & 2*, Oxford Psychologists Press.

[29] Stroop JR (1935) Studies of interference in serial verbal reactions., *Journal of experimental psychology* **18**, 6 643.

[30] De Renzi E, Faglioni P (1978) Normative data and screening power of a shortened version of the token test, *Cortex* **14**, 1 41–49.

[31] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, et al. (2011) The diagnosis of dementia due to alzheimers disease: Recommendations from the national institute on aging-alzheimers association workgroups on diagnostic guidelines for alzheimer's disease, *Alzheimer's & dementia* **7**, 3 263–269.

[32] Rascovsky K, Hodges JR, Knopman D, Mendez MF, Kramer JH, Neuhaus J, Van Swieten JC, Seelaar H, Dopper EG, Onyike CU, et al. (2011) Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia, *Brain* **134**, 9 2456–2477.

[33] Petersen RC, Caracciolo B, Brayne C, Gauthier S, Jelic V, Fratiglioni L (2014) Mild cognitive impairment: a concept in evolution, *Journal of internal medicine* **275**, 3 214–228.

[34] Extensible markup language (xml), http://www.w3.org/TR/REC-xml, Accessed on January 14, 2016.

[35] Salton G (1983) *Introduction to modern information retrieval*, McGraw-Hill.

[36] Bird S, Klein E, Loper E (2009) *Natural Language Processing with Python*, O'Reilly Media Inc.

[37] Perkins L, Whitworth A, Lesser R (1998) Conversing in dementia: A conversation analytic approach, *Journal of Neurolinguistics* **11** 33–53.

[38] Russell N, Norvig P (2003) *Artificial Intelligence a Modern Approach*, Prentice Hall.

[39] Smola A, Vishwanathan SVN (2008) *Introduction to Machine Learning*, Cambridge University Press.

[40] Bishop CM (2006) *Pattern Recognition and Machine Learning*, Springer.

[41] Pedregosa F, Varoquaux G (2011) Scikit-learn: Machine learning in python, *Journal of Machine Learning Research* **12** 2825–2830.

[42] Guyon I, Elisseeff A (2003) An introduction to variable and feature selection, *Journal of Machine Learning Research* **3** 1157–1182.

[43] National audit of memory clinics 2014, http://www.rcpsych.ac.uk/memoryclinicsaudit, Accessed on March 26, 2016.

[44] Menon R, Larner A (2011) Use of cognitive screening instruments in primary care: the impact of national dementia directives (NICE/-SCIE, National Dementia Strategy), *Family Practice* **28**, 3 272–276.

[45] Hessler J, Brnner M, Etgen T, Ander KH, Frstl H, Poppert H, Sander D, Bickel H (2014) Suitability of the 6CIT as a screening test for dementia in primary care patients, *Aging & Mental Health* **18**, 4 515–520.

[46] Boustani M, Callahan C, Unverzagt F, Austrom M, Perkins A, Fultz B, Hui S, Hendrie H (2005) Implementing a screening and diagnosis program for dementia in primary care, *Journal of General Internal Medicine* **20**, 7 572–577.

[47] Blackburn DJ, Wakefield S, Shanks MF, Harkness K, Reuber M, Venneri A (2014) Memory difficulties are not always a sign of incipient dementia: a review of the possible causes of loss of memory efficiency, *British Medical Bulletin* **112** 71–81.

[48] Jessen F, Wolfsgruber S, Wiese B, Bickel H, Mösch E, Kaduszkiewicz H, Pentzek M, Riedel-Heller SG, Luck T, Fuchs A, et al. (2014) Ad dementia risk in late mci, in early mci, and in subjective memory impairment, *Alzheimer's & Dementia* **10**, 1 76–83.

[49] Jessen F, Amariglio RE, Van Boxtel M, Breteler M, Ceccaldi M, Chételat G, Dubois B, Dufouil C, Ellis KA, Van Der Flier WM, et al. (2014) A conceptual framework for research on subjective cognitive decline in preclinical alzheimer's disease, *Alzheimer's & Dementia* **10**, 6 844–852.

[50] Paradise MB, Glozier NS, Naismith SL, Davenport TA, Hickie IB (2011) Subjective memory complaints, vascular risk factors and psychological distress in the middle-aged: a cross-sectional study, *BMC psychiatry* **11**, 1 1.