Automatic Detection and Rating of Dementia of Alzheimer Type through Lexical Analysis of Spontaneous Speech

Calvin Thomas, Vlado Kešelj, and Nick Cercone
Faculty of Computer Science
Dalhousie University

Kenneth Rockwood
Faculty of Medicine
Dalhousie University

Elissa Asp
English Department
Saint Mary's University

***Abstract*— Current methods of assessing dementia of Alzheimer type (DAT) in older adults involve structured interviews that attempt to capture the complex nature of deficits suffered. One of the most significant areas affected by the disease is the capacity for functional communication as linguistic skills break down. These methods often do note capture the true nature of language deficits in spontaneous speech. We address this issue by exploring novel automatic and objective methods for diagnosing patients through analysis of spontaneous speech.**

We detail several lexical approaches to the problem of detecting and rating DAT. The approaches explored rely on character n-gram-based techniques, shown recently to perform successfully in a different, but related task of automatic authorship attribution. We also explore the correlation of usage frequency of different parts of speech and DAT. We achieve a high 95% accuracy of detecting dementia when compared with a control group, and we achieve 70% accuracy in rating dementia in two classes, and 50% accuracy in rating dementia into four classes.

Our results show that purely computational solutions offer a viable alternative to standard approaches to diagnosing the level of impairment in patients. These results are significant step forward toward automatic and objective means to identifying early symptoms of DAT in older adults.

***Index Terms*— Automatic diagnostics, machine learning, natural language processing**

I. Introduction

Current methods of assessing dementia of Alzheimer type (DAT) in older adults involve structured interviews that attempt to capture the complex nature of deficits suffered. One of the most significant areas affected by the disease is the capacity for functional communication as linguistic skills break down. With this fact in mind, interviews are designed to test linguistic abilities, including confrontation naming [1], single word production [2] or word generation given context [3]. However, these methods sometimes fail to identify early symptoms observed by family members during normal conversation [4], and often fail to describe adequately the level of impairment in low scoring patients, unless similarities exist between performance during exams and in normal conversation [5]. In developing new tests, researchers should look for automatic and objective methods for use in rating dementia in patients through analysis of spontaneous speech that overcome the shortfalls of current methods [6]. Research advances in the areas of discourse analysis, language modeling and text classification may be applicable to this area and may lead to such progress.

In this paper, we detail several lexical approaches to the problem of detecting and rating DAT in patients from our corpus. The large corpus used in our research consists of transcripts from the Atlantic Canada Alzheimer's Disease Investigation of Expectations (ACADIE) study of the drug donepezil [7]. The goal of this research is to explore whether automatic techniques based on the analysis of spontaneous speech can provide objective measures of dementia levels in AD patients. It is our hope that improvements in automatic techniques will extend what is understood about the effects of dementia in Alzheimer's patients and the breakdown of language faculties.

The research discussed in this paper includes natural language processing and machine learning techniques that were applied to the problem of rating DAT in older adults. This interdisciplinary area brings opportunities for novel research to be conducted with generic text classification algorithms. Also explored are novel extensions to existing techniques that were developed to address specific qualities inherent to the corpus analyzed.

In short, we found that purely computational solutions offer a viable alternative to standard approaches to diagnosing the level of impairment in patients. Although more work needs to be done to improve the accuracy of these methods, these results are significant step forward towards automatic and objective means to identifying early symptoms of DAT in older adults.

II. Background and Related Work

Dementia of Alzheimer type. A significant component of the dementia of Alzheimer type (DAT) that accompanies Alzheimer's disease (AD) is aphasia, a loss of written and oral communicative ability [8], [9]. Symptoms of aphasia include breakdowns in semantic processing, shallow vocabularies and word-finding difficulties leading to the deterioration of spontaneous speech [10]. This deterioration begins early in the onset of the disease and is often observed by family members during conversational situations [4]. Further, recent studies of oral and written spelling have shown marked differences in language ability between AD patients and healthy older adults [11], [12].

For example, Ronald Reagan, former president of the United States, exhibited signs of AD from the outset of his presidency. Reagan's speeches suffered from word-finding difficulties, inappropriate phrases and uncorrected sentences that were obvious signs of his deterioration, but the fact that he had AD was not released until 1994 [13].

Current methods of assessing DAT levels in patients involve structured interviews that attempt to capture the breakdown of communicative capacity by testing specific linguistic abilities, including confrontation naming [1], single word production [2] or word generation given context [3].

However, these methods sometimes fail to identify early symptoms observed by family members during normal conversation [4], and often fail to describe adequately the level of impairment in low scoring patients, unless similarities exist between performance during exams and in normal conversation [5].

Mini-Mental State Exam. The Mini-Mental State Exam (MMSE) is a cognitive grading scale used in the assessment of patients first described by Folstein et al. [14] in 1975. This test addressed a need for a relatively short screening exam that could be used to reliably identify cognitive impairment in a clinical setting. Here, "mini" refers to the fact that this exam concentrates only on the cognitive impairment of mental function and excludes mental deficits covered by comprehensive exams, including mood and abnormal mental functions [14].

The MMSE involves a patient responding to 17 questions that cover a wide range of cognitive domains: orientation, registration, short-term memory, attention, calculation, visuospatial skills and praxis. Testing of the areas described above is divided into two sections; the first requires verbal responses to orientation, memory, and attention questions. The second section requires reading and writing and covers ability to name, follow verbal and written commands, write a sentence, and drawing intersecting pentagons. Testing time varies according to impairment level ranging between 5 and 10 minutes and can be administered by clinicians, nurses, psychologists, paramedical staff and lay interviewers, with limited training.

Since the introduction of the MMSE, this test has been widely used in clinical applications as an aid to diagnosis and in monitoring the progression of the dementia in individuals. The exam is also standardly used in the clinical and therapeutic research community as a basis for discretizing populations into normal, mild, moderate and severe dementia levels according to the DSM-IV [8]. Less standard, however, is the selection of boundary points in a community setting, since performance has been linked to level of education and other issues that may be characteristic of the population. With that said, "a variety of cutpoints have been suggested over the years, with 17/18 for clear-cut cases, 21/22, 23/24 and even 25/26" [15].

Verbal Picture Descriptions. Verbal picture descriptions can be used to assess the level of cognitive impairment and "are among the most sensitive measures for assessing spontaneous speech in AD" [10]. In these exams, the patient is supplied with a simple or complex line drawing that he or she must verbally describe. These narratives are recorded on tape and later analyzed according to a variety of speech attributes including articulation, grammar, phrase length, paraphasias, word-finding difficulties, themes and information content. While simple pictures may be useful in identifying patients with moderate deficiencies, more complex drawings may be helpful for screening patients with mild dementia [10], [13].

III. A Lexical Approach

Research in the area of automatic dementia detection in Alzheimer's patients has been quite limited, with few results found in a search of the literature [6], [16]. Bucks et al. [6] conducted a small study with 24 individuals: 8 patients and 16 healthy controls. The authors collected 8 lexical statistics over the first 1000 words of spontaneous speech during interviews, namely noun (N), pronoun (P), adjective (A) and verb (V) rates, type-token ratio (TTR), Brunét's Index (W), Honoré's Statistic (R) and the Clause-like Semantic Unit (CSU) rate. The results showed that the stylometric attributes had sufficient discriminating power in distinguishing between the language models of AD sufferers and control subjects.

N-rate, P-rate, A-rate and V-rate are the average rate of occurrence for each respective part-of-speech (POS) category. These measures capture the lexical distribution of the spoken words and were selected heuristically. Bucks et al. found that AD patients had "higher mean P-rate, A-rate, V-rate scores, but lower N-rate scores compared with normal older controls" [6].

The next three statistical attributes were selected to capture the lexical richness of the participant's speech. TTR is the ratio of the total vocabulary V to the overall text length N and is sensitive to the length of text collected. This measure

Attribute	Description	Comment
A-rate	Adjective rate	… on the *high* mountain …
N-rate	Noun rate	I went to my *house* …
P-rate	Pronoun rate	*He* came with *us* …
V-rate	Verb rate	Dave *likes* pizza …
TTR	Type token ratio	
W	Brunét's Index	
R	Honoré's Statistic	
CSU	Clause-like semantic unit	

TABLE I
ATTRIBUTE SET DESCRIBED IN BUCKS ET AL. [6]

Algorithm 1 *Profile_dissimilarity*($profile_1$, $profile_2$)

```
$sum \leftarrow 0$
**for all** n-grams $x$ contained in $profile_1$ and $profile_2$ **do**
    $f_1 \leftarrow$ frequency of $x$ in $profile_1$
    $f_2 \leftarrow$ frequency of $x$ in $profile_2$
    $sum \leftarrow sum + \left(\frac{2\cdot[f_1(x)-f_2(x)]}{f_1(x)+f_2(x)}\right)^2$
**Return** $sum$
```

is calculated as

$$TTR = \frac{V}{N} \tag{1}$$

where higher values are associated with a broader vocabulary. Brunét's Index W is a length insensitive version of TTR calculated using the following equation:

$$W = N^{V^{-0.165}} \tag{2}$$

The resulting value W typically ranges between 10 and 20, with richer speech producing lower values [17]. Honoré's Statistic R is also insensitive to length and is calculated as

$$R = \frac{100 \log N}{1 - \frac{V_1}{V}} \tag{3}$$

where V_1 is the number of words in the vocabulary only spoken once. Higher values of R indicate a richer vocabulary [18]. The CSU rate is a "measure of semantic cohesion in phrases … and characterizes the participant's ability to form noun and verb phrases and gives an indication of the flow of speech" [19]. To calculate this value, the corpus must first be hand-tagged according to a set of 13 rules that identify cohesion boundaries in phrases. The CSU rate is the average number of units found per 100 words. Patients suffering from dysphasia find it difficult to formulate long phrases leading to higher CSU rates than in normal speakers, making this variable "the most important discriminator between normal and dysphasic speech" [16]. Bucks et al. [6] confirmed that AD patients use less rich speech vocabulary according to the three lexical richness measures TTR, W and R. However, significant differences in CSU rates between AD patients and controls were not found in the data. Table I gives a summary of the attributes detailed above.

Common N-Grams (CNG) approach. The Common N-Grams (CNG) approach to authorship attribution uses character n-grams to model consistencies in author style. Traditional n-gram language models intuitively treat documents as a sequence of words and rely on word n-grams to capture consistencies with state-of-the-art performance [20]. However, several difficulties arise when working with word based methods, including language dependencies explicitly built into the model, word segmentation concerns and sparsity of data due to the large vocabulary. Overcoming these obstacles are particularly difficult when dealing with Asian languages such as Chinese or Japanese that do not have explicit word boundaries. By using byte-level n-grams the authors dramatically reduce the vocabulary, clearly define boundaries between units and do not make use of any language dependent information, including word boundaries, character case, white-space characters or punctuation [21]. However, due to their frequency and consistency of use by authors, white-space and punctuation characters implicitly play a significant role in classifier performance.

Author models are modeled by CNG profiles that are defined as "a set of the L most frequent n-grams with their normalized frequencies generated from training data" [21] and, hence, the two parameters of importance to the CNG method are n-gram size n and the profile length L. Due to the fixed and small vocabulary of ASCII characters used, the CNG method does not suffer from the sparse data problems of word n-gram approaches at low values of n. To be sure, the work in [21] indicates that values for $n \leq 8$ may be employed before computational limitations and performance decreases are encountered. This point contrasts with word-based approaches which are computationally feasible with values of n up to 3 or 4 [20]. The profile length L limits the number of n-grams considered during the similarity calculation and serves to keep profiles small when large values of n are used. Small profile lengths not only improve computational performance but also reduce model overfitting. This was supported by the fact that pruning threshold L was shown to improve accuracy with optimal values lying between 1000 and 5000 n-grams [21].

Classification via Common Word Frequencies. Using common word frequencies as style markers has be studied extensively by Burrows [22], [23], [24] and further investigated by Stamatatos et al. [25]. Both of these approaches focused on using the most frequent words in a text corpus as style markers. The primary difference between these two approaches is the training corpus from which these style

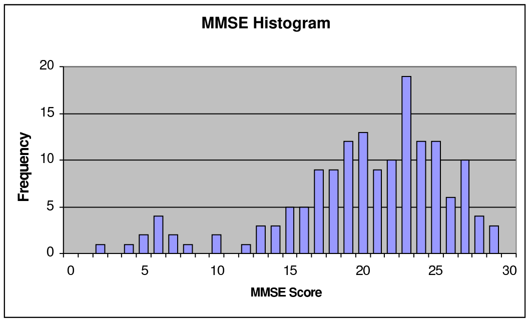

Fig. 1. Histogram of MMSE score

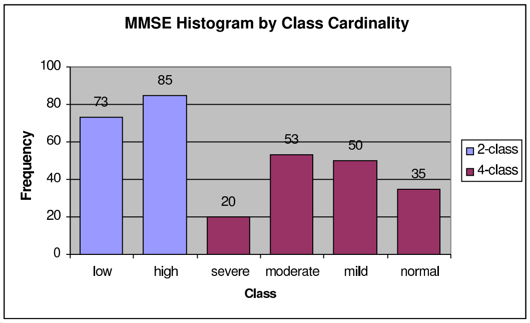

Fig. 2. Histogram of MMSE-based classes

markers were selected. Burrows argues for frequent terms that are selected from the target corpus itself and has shown effective classification results over a wide variety of literature domains [23], [24]. Stamatatos et al. [25] improved on previous results by extracting these style markers from the British National Corpus rather than the target corpus itself.

IV. Problem, Data and Solution

The research in this paper explores several approaches to the problem of automatically diagnosing the dementia level of Alzheimer's patients through analysis of spontaneous speech captured in a transcript. Each of these approaches assume that recognizable language artifacts, which are a function of the dementia level in patients, exist. Further, we are interested in attributes that can be extracted automatically from patient transcripts and can be used to reliably and consistently model the dementia level of AD patients.

ACADIE Dataset. The dataset used during analysis and experimentation contains the language spoken by 95 patients in 189 Goal Attainment Scaling interviews between field researchers, Alzheimer patients, and care-givers, compiled within the Atlantic Canada Alzheimer's Disease Investigation of Expectations (ACADIE) study of donepezil [7]. The dataset includes two interviews per patient with interviews conducted at assessment visits 12 weeks apart to examine the effects of the drugs administered during the interim. Interviews were conducted at six sites across Atlantic Canada [7].

MMSE scores are provided with the interview transcripts, with discretized scores in the ranges 0–15, 16–20, 21–24, and 25–30, according to [14].

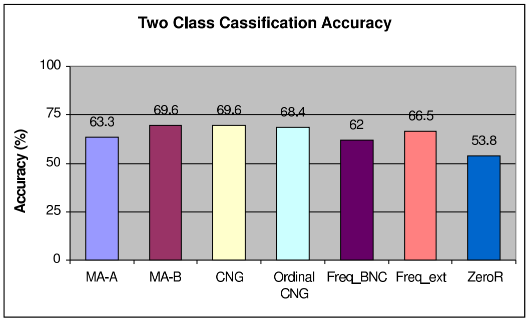

Fig. 3. Summary of accuracy on two-class task

V. Summary of the Results

Each of the figures in this section gives the classification performance in terms of maximum accuracy obtained for each explored approach on a specific classification task. Importantly, also included in each chart are the results from a naïve ZeroR rule-based classifier, which predicts the modal class during training for test instances. Overall, these results show that intelligent machine learning approaches performed better on the corpus than the naïve baseline of weighted random guessing. This indicates that pairing spontaneous speech data with machine learning techniques is a viable approach to the task of predicting dementia levels. Further, the results suggest improvements in classification accuracy are obtained by breaking large lexical categories into its smaller constituents by including modifier relationships.

Figure 3 illustrates the classification accuracies of the explored methods on the two class prediction task. In this task the classification algorithm must label test instances as low or $high$ scoring on the MMSE scale. Low indicates either a severe or moderate level of DAT impairment, while $high$ indicates that the patient should be placed in the $mild$ or $normal$ dementia classes. The ZeroR rule-based classifier produced a baseline accuracy of 53.8% for this task. From the other classifiers explored, an accuracy range of 62.0% to 69.6% was observed. On is task, the best accuracy was shared by MA_B and CNG at 69.6%, while trailing close behind was the ordinal CNG method with an accuracy of 68.4%.

The second classification task required the algorithm to predict one of four class labels for a test instance: $severe$,

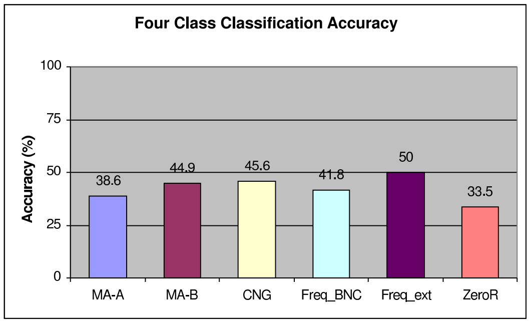

Fig. 4. Summary of accuracy on four-class task

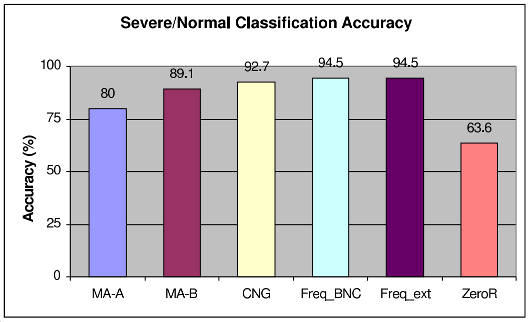

Fig. 5. Summary of accuracy on severe/normal task

$moderate$, $mild$ or $normal$. The results from this task are shown in Figure 4. On this task, a baseline accuracy of 33.5% was set forth by the ZeroR classifier, and a range of 38.6% to 50.0% was observed. The highest accuracy was achieved by classifiers using the $Freq_{ext}$ attribute selection method at 50.0%. The next best classifier was standard CNG with 45.6%, closely followed by MA_B with 44.9%.

Figure 5 compares the prediction accuracy for algorithms on a third task. This task involved predicting class labels for instances from the severe and normal groups only. The naïve baseline method produced an accuracy of 63.6% on this task. All of the intelligent methods examined in these experiments produced significantly higher classification accuracies with a range of 80.0% to 94.5%. Again, on this task the most accurate classifier was built over an attribute set consisting of frequent word ratios. Interestingly, both the $Freq_{BNC}$ and $Freq_{ext}$ produced the same classification accuracy at 94.5%. One other approach produced an accuracy above 90%, namely CNG at 92.7%. A particularly noteworthy observation is that the MA_B attribute set beat out MA_A by 9.1% on this task.

Figure 6 contains results from the mild/normal classification task. This task requires the algorithm to label test instances from $mild$ and $normal$ groups only. A baseline accuracy of 58.8% was posted for this task by the ZeroR classifier. The observed accuracy range for the other methods was between 60.0% and 75.3%. The MA_A attribute set performed the worst here and was only narrowly more accurate than the baseline. The best classification accuracy was achieved by the $Freq_{ext}$ attribute selection method at 75.3%. The next closest method in terms of classification accuracy was the other frequent words based method at 74.1%, a mere 1.2% behind.

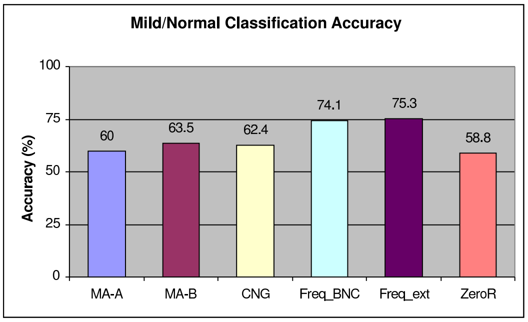

Fig. 6. Summary of accuracy on mild/normal task

VI. Conclusions

The thrust of this work was to examine the potential use of natural language processing and machine learning techniques in the diagnosis of dementia of Alzheimer type (DAT) in older adults. Framing this problem as a text classification task, we present several viable approaches based on mature algorithms and implementations. The main contributions are:

- a detailed statistical analysis of the lexical features exhibited in the spontaneous speech of older adults with Alzheimer's disease,
- novel application of several machine learning and natural language processing techniques in rating DAT,
- a novel classification algorithm in Ordinal CNG, and
- positive results in detecting DAT through an extensive exploration of classification methods.

1) Lexical analysis: A detailed statistical analysis was conducted on transcripts of spontaneous conversational speech collected from Alzheimer's patients. Analysis of spontaneous speech has the potential of offering many clues to the ties between linguistic ability and the extent of DAT. We chose to approach attribute selection from a statistical standpoint rather than rely on heuristics as in Bucks et al. [6]. We also believed that the detail of the Connexor part-of-speech tagger (POS) should be exploited to narrow the lexical categories analyzed. Our experiments confirmed the validity of our assumptions leading to higher accuracies and a better understanding of the data. During our lexical analysis of the data we found that closed class words were particularly helpful in predicting the level of language deficit in patients. Additionally, we found that lexical richness measures were not powerful discriminators for our purposes.

2) Novel application: Applying the CNG algorithm, which was originally developed for authorship attribution, to our DAT classification problem showed that the algorithm is robust with respect to application. The standard algorithm was applied without modification and achieved some of the most accurate results observed. This robustness is due to the byte-level n-grams used to construct the class profiles.

During our lexical analysis of the data we found that closed class words were helpful in predicting the level of language deficit in patients. Naturally, this lead us to examine in more detail these classes of words to determine if deeper relationships exist between the statistics and the observed effect in patients. Previous research had been done in the field of text classification where commonly used words were used as style markers. Our experiments showed that the novel approach to detecting deficit and novel application for these generic text classification algorithms were well suited for each other producing some of the most accurate models.

3) Algorithm extension: In addition to the standard CNG algorithm, an ordinal CNG extension was developed and tested. This algorithm was designed to take advantage of a natural ordering of classes, leveraging the training instances within the extreme groups. Our results showed that classification accuracy was not affected by the exclusion of $moderate$ and $mild$ training instances. This observation leads us to believe that our method effectively generates models using fewer training instances, but with better discriminating characteristics.

4) Positive results: The positive results reported in this work were arrived at after an extensive exploration of classification methods. This research showed that several standard classification algorithms could be used to produce classification accuracies significantly higher than our naïve rule-based classifier that always selects the modal class.

References

[1] J. Hodges, D. Salmon, and N. Butters, "The nature of the naming deficit in Alzheimer's and Huntington's disease," *Brain*, vol. 114, pp. 1547–1558, 1991.

[2] A. Martin and P. Fedio, "Word production and comprehension in Alzheimer's disease: the breakdown of semantic knowledge," *Brain and Language*, vol. 35, pp. 394–397, 1983.

[3] L. Phillips, S. D. Sala, and C. Trivelli, "Fluency deficits in patients with Alzheimer's disease and frontal lobe lesions," *European Journal of Neurology*, vol. 3, pp. 102–108, 1996.

[4] C. Crockford and R. Lesser, "Assessing functional communication in aphasia: Clinical utility and time demands of three mehods," *European Journal of Disorders of Communication*, vol. 29, pp. 165–182, 1994.

[5] S. Sabat, "Language function in Alzheimer's disease: a critical review of selected literature," *Language and Communication*, vol. 14, pp. 331–351, 1994.

[6] R. Bucks, S. Singh, J.M., Cuerden, and G. Wilcock, "Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analyzing lexical performance," *Aphasiology*, vol. 14, no. 1, pp. 71–91, 2000.

[7] K. Rockwood, J. Graham, and S. Fay, "Goal setting and attainment in Alzheimer's disease patients treated with donepezil," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 73, pp. 500–507, 2002.

[8] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed., Washington, DC, 1994.

[9] J. Cummings, F. Benson, M. Hill, and S. Read, "Aphasia in dementia of the Alzheimer type," *Neurology*, vol. 35, pp. 394–397, 1985.

[10] K. Forbes, A. Venneri, and M. Shanks, "Distinct patterns of spontaneous speech deterioration: an early predictor of Alzheimer's disease," *Brain and Cognition*, vol. 48(2-3), pp. 356–61, 2002.

[11] S. Pestell, M. Shanks, J. Warrington, and A. Venneri, "Quality of spelling breakdown in Alzheimer's disease is independent of disease progression," *Journal of Clinical and Experimental Neuropsychology*, vol. 22, pp. 599–612, 2000.

[12] H. Platel, J. Lambert, F. Eustache, B. Cadet, M. Dary, F. Viader, and B. Lechevalier, "Characterstics and evolution of writing impairment in Alzheimer's disease," *Journal of Clinical and Experimental Neuropsychology*, vol. 22, pp. 599–612, 1993.

[13] A. Venneri, O. Turnbull, and S. Della Salla, "The taxonomic perspective: the neuropsychological diagnosis of dementia," *Revue Europeenne de Psychologie Apllique*, vol. 46, pp. 81–86, 1996.

[14] M. Folstein, S. Folstein, and P. McHugh, "Mini-mental state. a practical method for grading the cognitive state of patients for the clinician," *Journal of Psychiatric Research*, vol. 12, pp. 189–198, 1975.

[15] C. Brayne, "The mini-mental state examination, will we be using it in 2001?" *International Journal of Geriatric Psychiatry*, vol. 13, pp. 285–294, 1998.

[16] D. Holmes and S. Singh, "A stylometric analysis of conversational speech of aphasic patients," *Literary and Linguistic Computing*, vol. 11, pp. 45–60, 1996.

[17] E. Brunét, "Le vocabulaire de jean giraudoux," *Structure et Evolution*, 1978.

[18] A. Honoré, "Some simple measures of richness of vocabulary," *Association of Literary and Linguistic Computing Bulletin*, vol. 7, pp. 172–177, 1979.

[19] S. Singh, "Computational analysis of conversational speech in dysphasic patients," Ph.D. dissertation, University of the West of England, UK, 1996.

[20] F. Peng, D. Schuurmans, V. Keselj, and S. Wang, "Automated authorship attribution with character level language models," in *Proceedings 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, 2003.

[21] V. Keselj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution," in *Proceedings of Pacific Association for Computational Linguistics (PACLING'03)*, 2003.

[22] J. Burrows, "Word-patterns and story-shapes: The statistical analysis of narrative style," *Literary and Linguistic Computing*, vol. 2, no. 2, pp. 61–70, 1987.

[23] ——, "Not unless you ask nicely: The interpretative nexus between analysis and information," *Literary and Linguistic Computing*, vol. 7, no. 2, pp. 91–109, 1992.

[24] ——, "'Delta': a measure of stylistic difference and a guid to likely authorship," *Literary and Linguistic Computing*, vol. 17, no. 3, pp. 267–287, 2002.

[25] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Text genre detection using common word frequencies," in *Proceedings of 18th International Conference on Computational Linguistics (COLING2000)*, vol. 2, 2000, pp. 808–814.