Automatic Detection of Linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: A Computational Linguistics analysis

Vassiliki Rentoumi, George Paliouras
Inst. of Informatics and Telecommunications,
NCSR "Demokritos", Athens, Greece

Dimitra Arfani, Katerina Fragkopoulou,
Spyridoula Varlokosta,
Department of Linguistics, National and Kapodistrian University of Athens (UoA), Greece

Eva Danasi
Athens Association of Alzheimer's Disease and Related Disorders, Athens, Greece

Spyros Papadatos
Department of Automating systems,electronics & computer sciences, Technological Educational Institute of Central Greece

***Abstract*-In the present study, we analyzed written samples obtained from Greek native speakers diagnosed with Alzheimer's in mild and moderate stages and from age-matched cognitively normal controls (NC). We adopted a computational approach for the comparison of morpho-syntactic complexity and lexical variety in the samples. We used text classification approaches to assign the samples to one of the two groups. The classifiers were tested using various features: morpho-syntactic and lexical characteristics. The proposed method excels in discerning AD patients in mild and moderate stages from NC leading to the in-depth understanding of language deficits.**

Keywords--machine learning, neurolinguistics, cognitive linguistics, text classification

I. INTRODUCTION

Degenerative conditions, such as Alzheimer's disease (henceforth AD) are commonly associated with deficits across a range of subcomponents of linguistic competence.Although both AD and other types of dementia are associated with changes in spoken and written language, these changes have not been extensively examined or compared. Memory impairment implies that the vocabulary of patients with dementia is poorer and simpler than that of healthy subjects and more incoherent. Language expression shows that neuronal brain activity in the area of Broca and Wernicke is reduced, so that words often do not make sense and complement the information lost with the damaged regions neuronal cells.

In the present study, we analysed samples obtained from Greek native speakers diagnosed with Alzheimer's in mild and moderate stages and from age-matched cognitively normal controls (henceforth NC). Initially, we looked for differences in language between AD patients and NC by employing various quantitative methods of evaluation. We also searched for the most important sources and criteria of linguistic variation between these groups. Most importantly, we tried to identify linguistic structures that diverge from the language norm. These indicators of language divergence could be used to help the early diagnosis of AD and other dementias by facilitating the clinical process of doctors.

As far as the linguistic features of AD are concerned, syntactic complexity and lexical variation cues were investigated to identify AD-induced changes. We adopted a computational approach for the comparison of morpho-syntactic complexity and lexical variation in written texts produced by AD patients and by cognitively NC. Analysis has been obtained employing the Alzheimer's detector[1]. Further, we used text classification approaches to assign the samples to one of the two groups. The classifiers were tested using features such as morpho-syntactic complexity and lexical characteristics. The proposed method succeeds in finding linguistic characteristics which differentiate AD patients in mild stages from NC leading to in depth understanding of language deficits.

Further, going beyond other approaches, which focus on analyzing syntax and lexical features in spoken language of AD patients, through the proposed approach we intend to verify our conjectures that lexical and syntactic complexity domains are also affected in written language due to the early onset of AD. Such deficits reflected in written language can also help us in discriminating AD

[1] The Alzheimer's detector is a textual analysis program designed to parse Greek language samples and to provide a range of syntactic complexity and lexical variation measures.

978-1-5386-1264-4/17/$31.00 ©2017 IEEE

patients from NC employing written text samples. Such a verification,pursued for the first time in Greek data,could pave the way towards the implementation of novel cross-linguistic diagnostic tools which contribute to the timely diagnosis of the early AD. The timely diagnosis of early AD can contribute to the assessment in the language in correlation to behavior and cognition in this population.

The presented work comprises an effort to enhance interdisciplinary research among the domains of cognitive sciences, linguistics and infocommunications proposing a novel cognitive engineering application. We provide a computational foundation for the assessment of cognitive function and consequent intervention with the goal to preserve and optimize human cognitionvia an efficient human computer interface.

In what follows we initially present the state of the art in Section 2 while in Section 3 we describe the proposed methodology. In Section 4, results derived from the application of the proposed methodology are presented. Finally, in Section 5 the paper is concluded with a short discussion and with sketching plans and future directions.

II. Existing Work on Computational Language Analysis as a Diagnostic Means of Dementia

Over the last five years, several researchers have used computational methods to study spontaneous or connected speech to find language features that can distinguish the speech of individuals with AD, MCI or other types of dementia from the speech of healthy adults, aiming at early detection of neurological disorders through speech analysis. Machine learning methods (henceforth ML) were used to discriminate the speech of patients with semantic dementia from healthy elderly, based on lexical characteristics in transcribed narratives [1]. According to results, the vocabulary of semantic dementia and AD is characterised by more generic terms (e.g., "thing", "something") and components of metanarrative statements (e.g. "know", "remember"), while a number of low frequency content words (e.g. "sailing" and "pail") occured only in the vocabulary of healthy elderly [1].

Similarly, in [2] ML methods were used to automatically extract features from digital samples of connected speech to distinguish among different subtypes of primary progressive aphasia .

Concerning the language deficits in AD, it has been found that patients with AD make more lexical errors (word-finding difficulties, repetitions, phonemic paraphasias and semantic substitutions) than the control group and use less syntactically complex sentences [3]. Moreover, characteristics of language complexity have been measured (e.g. words per clause, content density and prosodic features like the length of pauses, total pause and phonation time) and it has been proposed that a combination of computational methods can differentiate patients with MCI and NC with respect to language complexity [4].

Further, greater deficits in syntactic complexity and language variation have been recorded in the spoken language of AD patients with vascular load compared to AD patients [5]. AD patients have also difficulty in producing syntactically complex sentences, especially in reduced sentences and they manifest significant differences in the production of predicates, in repetitions, word replacement and they produce incomplete words in comparison to NC [6]. ML has been applied to study linguistic features in AD, such as semantic substitutions, syntactic complexity, length of noun, verb and adjectival phrases, parts of speech and their frequency, vocabulary richness, information content, repetitiveness, phonological errors that can be extracted automatically from digital samples of connected speech [7]. A semantic impairment with increased use of repetitions and pronouns, and reduced use of lexical variety was detected, as well as a syntactic deficit in the production of auxiliary verbs, gerunds and participles [7].

There has been comparatively little work done on analyzing written language spontaneously generated by people with AD. Such an investigation can offer invaluable insights regarding language status and patterns characterizing patients with AD as written language bristles with more carefuland more complete linguistics structures than the spoken language which is more spontaneous and therefore more careless.

III. METHODS PROPOSED

A.Patients Data

Data from 60 subjects have been used for the current study, in the 30 of whom a pathological diagnosis of AD has been made, while the remaining 30 have been diagnosed as NC. Participants in the AD group 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 both groups were matched as closely as possible for age, gender distribution and years of education (see below table I). There were not any statistical significant differences between groups regarding age, education, gender, while MMSE scores between the two groups showed statistically significant differences ($p<0.05$).

TABLE I. DEMOGRAPHICS

Demographics	***AD (n=30) Mean***	***Controls (n=30) Mean***	
Age (years)	66.48	68,03	n.s
Education (years)	12	13,93	n.s
Gender (male: female)	13: 17	16:14	n.s
MMSE scores	22.68	28.26	*

*Table I: AD and NC groups' comparison of demographics and MMSE scores. Comparisons of groups have been computed using Independent Samples T-test and Chi-Square for gender ratios. Means (M) of demographic data. n.s = not significant, *p<0.05*

B.Language Data

The data for the present study consisted of a total of 60 written discourse samples obtained from the two groups. These written samples have been obtained employing the

TABLE II- FEATURES OF VOCABULARY VARIATION AND SYNTACTIC COMPLEXITY

Lexical Variation measures	1. Lexical word variation	LV	Number of Lexical word types/ number of lexical words
	2. Bi Logarithmic type token ratio	Log TTR	Log_x word types/ Log_x total words (tokens)
	3. Noun variation	NV	Number of noun types/ number of lexical words
	4.Adjective variation	ADJV	Number of adjective types/ number of lexical words
	5.Modifier variation	MODV	Number of adjective types + adverb types/ number of lexical words
	6.Adverb variation	ADVV	Number of adverb types/ number of lexical words
	7.Corrected verb variation	CVV	Number of lexical verb types/ number of verbs x 2
	8.Verb variation	VV	Number of lexical verb types/number of lexical words
	9. Brunet	W	$N^{v-0.165}$ (N = number of word tokens ; V = vocabulary)
Syntactic complexity measures	10.Mean length of sentence	MLS	Number of words/ number of sentences
	11.Mean number of noun phrases	MNP	Number of noun phrases / Number of all sentences of each text

Cookie Theft picture description task which comprises part of the Boston Diagnostic Aphasia Examination and is also component of the CAMCOG [8]. To our knowledge it is the first time that computational methods are developed to detect early signs of dementia from the analysis of Greek written language and most importantly employing the written descriptions of the Cookie theft picture. The latter task has been usually administered to people to elicit spoken or written language.

In our case the picture has been shown to the groups of people and they have been asked to produce a written description of the picture. This task is described as following: a picture is given to the subjects and the researcher asks the subject to describe the picture in as far as possible complete and adequate way. The examiner tries to elicit an unbiased, qualitative and objective written sample. The examiner does not interfere during the writing of the subject, but only when help or clarification is needed. If the examiner does not succeed to elicit a satisfactory sample of data, he/she proceeds to further elicitation questions, such as "And what else can you see in the picture? or "Have you described all the detail of the picture?" etc.

C. Written Samples' Analysis

The aim of the analysis was to classify every written language sample correctly with respect to its subgroup (AD versus NC) of origin. The analysis is articulated in two consecutive steps. The output of the first stage is providing the input for the second.

In the first stage values associated with 10 features(defined in Table II) have been extracted from each written sample. Feature extraction is performed using the Alzheimer's detector which computes a range of lexical variation and syntactic complexity measuresin syntactically parsed Greek texts.The syntactic analysis of samples is performed using the Part of Speech (PoS) tagger [10] and the NP chunker [11] for Greek which are both part of the Ellogon [12] service[2], a tool for Natural Language Processing. The Alzheimer's detector has been inspired by Lu's L2 Syntactic and Lexical complexity analyzers [17],[18] that are designed to parse English texts. Lu's analyzers have been previously employed for the analysis of English speech transcripts produced by patients ofvarious forms of Alzheimer's [5].

In the second and final stage of the process the extracted features were employed to 'train' classifiers to assign samples to one of the groups under comparison. ML classification is implemented using Waikato Environment for Knowledge Analysis (WEKA)[13] http://www.cs.waikato.ac.nz/ml/weka. Methodology for both stages is described in the following two sections.

D. Feature Extraction

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

Lexical variation features comprise measures of the vocabulary range in each written sample group, as estimated using lexical word variation (LV)—the ratio of the number of lexical word types to the total number of lexical words in a text and the bilogarithmic type/token ratio (Log TTR), higher values of which indicate greater lexical variation [14]. Features 3-8 are more fine-grained variants of LV. They share the same denominator (the number of lexical words), but employing counts of nouns, verbs, adverb, adjectives and modifier types as numerators [15].A related index is Brunet's W, lower values of which imply a higher number of distinct word types, and thus a richer vocabulary.

Further features 10-11 are accounted for syntactic complexity which can be quantified by means of the number of immediate constituents of a syntactic construction. Syntactic complexity shows the degree to which grammatical structures vary in writing. In the current

[2]http://www.ellogon.org/clarin-ellogon-services/process/system/np_tagger.tcl

approach we focus on computing the length of sentences and the amount ofnoun phrases per written sample.

E. Machine Learning Classification

We employed a ML classification approach to predict the group (AD or NC) to which each participant's written sample belonged.

For each written sample, the corresponding values of the extracted features formed a feature vector representing each the written sample. More formally, each letter was represented by a feature vector of the form $(\vec{x}, y)$ where $\overrightarrow{x} \in X$is a feature vector consisting of a sequence of values $(x_1, \dots x_n)$, for each feature and a binary label$y \in \{e.g. AD, NC\}$.

A machine learning classifier is provided as input with a set of vector representations of written samples already assigned to a group (training data). The trained classifier is then given a new set of unseen written samples (test data), each of which it assigns to one of the groups using the model derived from the training data.

In our experimental setting, we employed two well-known classifiers for applying text classification namely the Naïve Bayes(NB) [13] and Support Vector Machine using the SMO (Sequential Minimal Optimization) [14]training algorithm.

The NB classifier is an implementation of Bayes' theorem. The term 'Naïve' derives from the fact that the classifier assumes the features it employs 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 [13].

The NB classifier learns estimates for the category-conditional probabilities and priors for each group (AD, NC) from the training data. At the classification stage, Bayes' theorem is used to assign a feature vector representing a written sample to the class that maximizes the posterior probability (i.e., the more likely class) [13]. The Sequential Minimal Optimization (SMO) [14] classifier is one way tosolve the support vector machines (SVM) optimization training problem. SMO uses heuristics to partition the training problem into smaller problems that can be solved analytically.

TABLE III: MICRO-AVERAGE CLASSIFICATION ACCURACY OF ML CLASSIFIERS VS BASELINE FOR COMPARISONS A & B

	Comparison	Naïve Bayes			SMO classifier			Baseline (ZeroR)			p-value[3]
		Correct	Incorrect	Accuracy	Correct	Incorrect	Accuracy	Correct	Incorrect	Accuracy	
A	AD (n[4]=30)	21	9	0.78	24	6	0.80	100	0	0.50	< 0.00001.
	NC (n=30)	4	26		6	24		100	0		
B	AD (n=100)	76	24	0.85	88	12	0.885	30	0	0.5	< 0.00001.
	NC (n=100)	6	94		11	89		30	0		

Doing so we avoid using a time-consuming in the training process. Both NB and SMO have been proven very efficient in text classification problems [15]. For our experiments, we employed NB through the WEKA package. Additionally, we used the SVM classifier, using the SMO WEKA implementation.

F. Evaluation Procedures

To evaluate classification procedures, we adopted a 10-fold cross validation approach. First, the written samples used for the classification task were randomly divided into 10 equally sized, randomly selected subsets, nine of which were used to train the classification algorithm. The remaining subset was used to test the classification algorithm. The classification step was repeated ten times, with a different subset used for testing purposes each time, and training on the remaining nine subsets.

[3]Independent samples t-test :Naïve Bayes, SMO vs. ZeroR baseline

[4]Number of text instances

IV. RESULTS

The NB and SMO classifiers have been used for comparisons A and B. The baseline condition was implemented using the ZeroR classifier (http://chem-eng.utoronto.ca/~datamining/dmc/zeror.htm) provided by WEKA, which predicts the majority category. Since it lacks predictability power, ZeroR is only useful for determining a baseline performance as a benchmark for other classification methods.

While for comparison A we employed our dataset consisting of 60 written samples, 30 from each category (AD, NC), in comparison B we created a synthetic dataset consisting of 200 samples, 100 from each category (AD, NC), to test the effectiveness of our method towards a larger data set. The SMOTE [16] algorithm was employed to create the synthetic sample of comparison B, by expanding the data set of the 30 samples of each category (AD, NC) providing an additional dataset of 70 written samples for each category. Thus, we ended up with a synthetic sample of 100 (30 real samples, 70 synthetic samples) sample data for each category (AD, NC).

Table 3 shows the confusion matrices which report the numbers of written samples in each comparison that were classified as True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) and the micro-average classification accuracy scores, for comparisons A and B. The micro-average accuracy is calculated using the TP, TN, FP and FN counts from each of the 10 folds, according to (1):

(1)

$$\frac{\sum_{i=1}^{n} TP_i + TN_i}{\sum_{i=1}^{n} TP_i + TN_i + FP_i + FN_i}$$

Independent Samples t-tests were conducted on the micro-average accuracies compared to the baseline condition, and were computed separately for each fold.

In the classification of written samples for both comparisons A and B, in both classification tasks NB and SMO significantly outperformed the baseline condition. Such a finding implied that the language of the AD group, by means of the features identified, is distinguishable from the language in the samples of the NC group. Further both comparisons argue in favor of the existence of an idiosyncratic language pattern characterizing the language of the AD group. This language pattern has been identified by means of lexical variation and syntactic complexity differences, verifying that the latter are very good discriminating factors when it comes to distinguish the language of AD and NC groups. Finally, the fact that the performance of the classifier stays stable when it is evaluated towards a bigger sample (Comparison B) argues in favor of the robustness of the proposed method.

V. CONCLUSION AND FUTURE STEPS

The potential of ML classification to classify the language of different clinical populations has been primarily investigated within the context of spoken language. This is the first time ML methods are evaluated towards written language data for the diagnosis of AD and, more specific it is the first time a research investigates Greek written samples produced by Greek native speakers for this purpose. Such an endeavor aims at identifying specific language idiosyncratic features bound to AD but also to confirm the presence of cross linguistic deficits caused by AD. Most importantly the latter finding can pave the way towards the creation of a cross linguistic diagnostic tool for AD.

In the present study a ML classification approach employing a combination of lexical variation and syntactic complexity features outperformed a baseline approach. Adopting a 10-fold cross validation approach we eliminate the possibility that the classifier over-fitted on the language idiosyncrasies of the training set which could yield poor generalization on unseen data. The high accuracies achieved in both comparisons imply that the classifiers' performance was high in all the 10 fold classifications tasks.

It is important to note that high accuracy can be further seen as a result of an appropriate combination of ML classifier with the features employed. Both classifiers employing the same set of features (features of syntactic complexity and lexical variations) noted a consistently high performance in all comparisons (A and B) outperforming the baseline condition in all cases. Such a finding argues in favor of the existence of systematic differences in lexical variation and syntactic complexity in groups under comparison.

The current approach verifies our primary research hypothesis that cognitive deficits of AD patients can be reflected in their written language and are evidenced in both lexical variety and syntactic complexity domains. In the future we intend to acquire a more extended data from patients that suffer from AD but also from other forms of dementia so that we can detect differences between a wider range of dementia pathologies and healthy elderly. Further we intend to enrich our feature set with additional syntactic complexity features as well as with word and character n-grams frequencies that have been shown to be effective in discriminating along clinical dimensions.

ACKNOWLEDGMENT

This work was partially supported by the EU H2020 programme, under grant agreement No 727658 (project IASIS).We would like to thank all the participants and their families who have provided the written consent to share the data for the purposes of the current research. Additionally we would like to cordially thank Dr. Magda Tsolaki from Alzheimer Hellas and Dr. Paraskevi Sakka from Athens Association of Alzheimer's Disease and Related Disorders who kindly agreed to help us in the sample collection processes. Finally we would like to thank staff of the Athens Association of Alzheimer's Disease and Related Disorders (Alzheimer's Center Panormou) who helped us in the data collection and neuropshycological assessment of participants.

REFERENCES

[1] P. Garrard, V. Rentoumi, B. Gesierich, B. Miller, and M.L., Gorno-Tempini, "Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse", Cortex Vol. 55, pp. 122-129, June 2014.

[2] K.C. Fraser, JA. Meltzer, NL Graham, C. Leonard, G. Hirst, SE Black, and E. Rochon, "Automated classification of primary progressive aphasia subtypes from narrative speech transcripts", Cortex, Vol. 55, pp. 43-60, June 2014.

[3] J.O. de Lira, K.Z. Ortiz, A.C. Campanha, P.H.F. Bertolucci, and T.S.C., Minett, "Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease", International Psychogeriatrics Vol. 23 no. 3, pp. 404-412, April 2011.

[4] B. Roark, M. Mitchell, J.P Hossom, K. Hollingshead, and J. Kaye. "Spoken language derived measures for detecting mild cognitive impairment", IEE Transactions on audio, speech and language processing Vol. 19 no. 7, pp. 2081-2090, 2011.

[5]Rentoumi, Vassiliki, LadanRaoufian, Samrah Ahmed, Celeste A. de Jager, and Peter Garrard. "Features and machine learning classificationof connectedspeech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology." Journal of Alzheimer's Disease 42, no. s3 (2014): S3-S17.

[6] S. Orimaye, J.S-M. Kong, K.J. Golden, C.P. Wong, and I.N., Soyiri. "Predicting probable Alzheimer's disease using linguistic deficits and biomarkers", BMC Bioinformatics Vol. 18 no 1, pp.1-13, 2017.

[7] K.C. Fraser, J.A. Meltzer, and F., Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech". Journal of Alzheimer's Disease Vol. 49 no. 2, pp.407-422, 2016.

[8] Roth, M. T. Y. M. E., et al. "CAMDEX. A standardised instrument for the diagnosis of mental disorder in the elderly with special reference to the early detection of dementia." The British journal of psychiatry 149.6 (1986): 698-709.

[9] Holstein, M. F., S. E. Holstein, and P. R. McHugh. "Mini-mental state. A practical method for grading the cognitive state of patients for the clinician." J Psychiatr Res 12.3 (1975): 189-198.

[10] Petasis, Georgios. "The Ellogon Pattern Engine: Context-free Grammars over Annotations." LREC. 2014.

[11] Petasis, G. E. O. R. G. I. O. S., et al. "Using machine learning techniques for part-of-speech tagging in the Greek language." Advances in Informatics. Singapore: World Scientific (2000): 273-81.

[12]Georgios Petasis, Vangelis Karkaletsis, Georgios Paliouras, Ion Androutsopoulos and Constantine D Spyropoulos. Ellogon: A New Text Engineering Platform. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002). 2002, 72–78.

[13] Hall, Mark, et al. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11.1 (2009): 10-18.

[14] Engber, Cheryl A. "The relationship of lexical proficiency to the quality of ESL compositions." Journal of second language writing 4.2 (1995): 139-155.

[15] McClure, Erica. "A comparison of lexical strategies in L1 and L2 written English narratives." Pragmatics and language learning 2 (1991): 141-154.

[13] A. McCallum, and K. Nigam, "A comparison of event models for Naive Bayes text classification", AAAI/ICML98, 1998 [Workshop on Learning for Text Categorization, pp. 41-48].

[14] J. Platt. "Fast Training of Support Vector Machines using Sequential Minimal Optimization". In B. Schoelkopf and C. Burges and A. Smola (Editors), Advances in Kernel Methods - Support Vector Learning. MIT Press (1998).

[15] Trivedi, M., et al. "Comparison of Text Classification Algorithms." International Journal of Engineering Research & Technology (IJERT) 4.02 (2015).

[16] Pears, Russel, Jacqui Finlay, and Andy M. Connor. "Synthetic Minority over-sampling technique (SMOTE) for predicting software build outcomes." arXiv preprint arXiv:1407.2330 (2014).

[17]Lu, Xiaofei. "Automatic analysis of syntactic complexity in second language writing." *International Journal of Corpus Linguistics* 15.4 (2010): 474-496.

[18] Lu, X., 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), pp.190-208.