Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 10 (2018) 260-268

Cognitive & Behavioral Assessment

Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task

Laura Hernández-Domínguez[a,*], Sylvie Ratté[a], Gerardo Sierra-Martínez[b], Andrés Roche-Bergua[c]

[a]*École de technologie supérieure, Université du Québec, Montreal, Quebec, Canada*
[b]*Engineering Institute, Universidad Nacional Autónoma de México (UNAM), Mexico City, Mexico*
[c]*Psychogeriatric Unit, Hospital Psiquiátrico Fray Bernardino Álvarez, Mexico City, Mexico*

Abstract

Introduction: We present a methodology to automatically evaluate the performance of patients during picture description tasks.

Methods: Transcriptions and audio recordings of the Cookie Theft picture description task were used. With 25 healthy elderly control (HC) samples and an information coverage measure, we automatically generated a population-specific referent. We then assessed 517 transcriptions (257 Alzheimer's disease [AD], 217 HC, and 43 mild cognitively impaired samples) according to their informativeness and pertinence against this referent. We extracted linguistic and phonetic metrics which previous literature correlated to early-stage AD. We trained two learners to distinguish HCs from cognitively impaired individuals.

Results: Our measures significantly ($P < .001$) correlated with the severity of the cognitive impairment and the Mini–Mental State Examination score. The classification sensitivity was 81% (area under the curve of receiver operating characteristics = 0.79) and 85% (area under the curve of receiver operating characteristics = 0.76) between HCs and AD and between HCs and AD and mild cognitively impaired, respectively.

Discussion: An automated assessment of a picture description task could assist clinicians in the detection of early signs of cognitive impairment and AD.

Keywords: Alzheimer's disease (AD); Mild cognitive impairment (MCI); Picture description task; Automatic assessment; Information coverage; Linguistic analysis; Phonetic features; Machine learning

1. Introduction and motivation

Multiple studies have assessed language functions as early markers of Alzheimer's disease (AD) [1]. Consequently, language is now widely accepted to be one of the first cognitive abilities affected by this dementia. Some of the most commonly used tests in clinical practice are Verbal Fluency by categories, Picture Description, the Boston Naming Test [2], and the Token Test [3], which measure expository speech, oral expression, and comprehension of commands, respectively [4].

This exploration of the changes in language functions derived from AD has attracted significant attention among scientists outside the field of medicine [5]. Researchers, especially those working in natural language processing, have proposed computer-based approaches for automatic and semiautomatic analysis of language in patients suffering from AD [6–13].

In this work, we propose a methodology to automatically describe patients' performance during a picture description task [14]. We selected this type of test because it elicits spontaneous speech from patients, allowing us to describe not only patients' ability to retrieve information from a visual stimulus but also some of their

The authors have declared that no conflict of interest exists.

*Corresponding author. Tel.: +1-514-431-1557.
E-mail address: laudobla@gmail.com

https://doi.org/10.1016/j.dadm.2018.02.004

linguistic characteristics. Our evaluation describes three aspects: the informativeness and pertinence of the description provided by the patient, some linguistic characteristics, such as vocabulary richness and general use of part-of-speech categories, and a phonetic overview.

1.1. Information coverage

One of the key objectives of a picture description task is to measure the amount and quality of the information that a patient can provide from a visual stimulus. Even early in the course of the disease, AD patients have been shown to provide less informative descriptions than cognitively intact elderly adults [15]. This measure is generally made by comparing the description provided by the patient to a list containing the main information content units (ICUs) of the image, namely, actors, objects, actions, and places. Over the years, several authors have come up with predefined lists of ICUs for the Cookie Theft picture description task [16–21]. However, one of the disadvantages of using predefined lists to evaluate elderly patients is that the list author does not necessarily have a similar education level, age, focus, cultural background, and interests as the target population. Also, different authors may come up with different lists, depending on their idiosyncrasies, their own observations, and what they may consider "important" from the picture.

1.1.1. Related computational works

Hakkani-Tür et al. [22] used a manually predefined list as a referent to automatically compare descriptions of the Western Aphasia Battery's Picnic Picture. The authors found a high correlation between the traditional manual assessment and their automated approach. However, the computer-based evaluation had trouble handling ICUs expressed in multiple ways.

Pakhomov et al. [23] used manual transcriptions of descriptions of the Cookie Theft picture to assess the performance of patients with frontotemporal lobar degeneration. They compiled a list of predefined ICUs based on Yorkston and Beukelman's study [16] and manually extended it to include lexical and morphological variants of words and phrases. One drawback of this method is that it entails the manual creation of a list that considers as many variants as possible for each ICU.

Fraser et al. [24] used a semiautomatic approach to automatically classify Alzheimer's patients and healthy elderly controls (HCs) by analyzing manual transcriptions of descriptions of the Cookie Theft picture in the Pitt Corpus [25]. As a referent, the authors used the predefined list proposed by Croisile et al. [19] and evaluated the frequency of key words used to name the ICUs in different ways. As in Pakhomov et al.'s work [23], manually considering all the ICUs and their linguistic variations is a time-consuming task.

Yancheva and Rudzicz [26] automatically extracted the main ICUs retrieved by elderly adults in the Pitt Corpus. The authors contrasted automatically extracted ICUs to a combination of several predefined lists of ICUs. They retrieved most of the human-selected ICUs. In addition, they found that some participants mentioned the object *apron*, a new ICU that none of the specialists had perceived before. They also observed that HCs were more prone than AD patients to mention this object in their descriptions.

The appreciation of the fact that a woman is wearing an apron while doing housework could be attributed to a generational and cultural perception of what the object *apron* represented to elderly participants taking the test back in the 1980s. Different remarks may be attributable to cultural differences. For example, a non-Caucasian-predominant population may remark on the fact that all the subjects in the Cookie Theft picture are blond. Hence, we consider that a fairer referent for comparison in this task should be constructed by healthy participants of the target population. As such, it would be possible to create referents that are adapted to specific populations from different generations, cultures, and educational and general socioeconomic backgrounds.

1.1.2. The coverage measure

We identify three important tasks for performing a computer-based evaluation of a picture description task:

1. Creating a population-adapted referent.
2. Evaluating the *informativeness* of descriptions: estimate how much of the information in the referent is being covered by the participant.
3. Evaluating the *pertinence* of utterances: determine how much of what the participant is saying is covered by the referent. Some participants, particularly those with AD, can drift off-topic. Although this situation is easily detected when performing a manual evaluation, it is a challenging task for an automated analysis.

With these tasks in mind, we selected the information coverage measure proposed by Velazquez [27]. He originally proposed the method for comparing the coverage of information in news articles, although it could be used in different contexts.

Velazquez proposes a methodology for creating a referent for evaluating the information coverage. One distinguishing feature of his measure is that it uses linguistic patterns that allow the consideration of the context. In addition, the measure allows a two-way analysis of the information coverage, from the referent by the subject of comparison and vice versa. These two measures would allow the estimation of informativeness and pertinence, respectively.

1.2. Linguistic characteristics

There is extensive literature covering the analysis of the linguistic characteristics of AD patients [6,7,24,28–34]. As part of our evaluation, we selected those that most authors

have found to correlate significantly with the disease and that could be used in picture description tasks (Table 1). In Section 2.3, we provide further information about the methodology and tools used for extracting these characteristics.

1.2.1. Part-of-speech distribution

We made an evaluation of the frequency and ratio of adjectives, conjunctions, nouns, prepositions, and verbs per 100 words. We also evaluated the frequency of auxiliary verbs and their ratio to the total number of verbs.

1.2.2. Vocabulary richness

Several measures have been explored to evaluate the richness of an author's language. These same measures can be used to evaluate the variability of the vocabulary of patients during a picture description task.

1.3. Phonetic analysis

Several authors [23,24,35–38] have found significant differences in the audio signals produced by AD patients as compared to cognitively intact elderly individuals. Mel-Frequency Cepstral Coefficients (MFCCs) are among the most used features for automatic speech analysis. Only the first 12 to 13 MFCCs are usually used because most of the information about the transfer function of the vocal tract is in the lower range of frequencies.

2. Methods

2.1. Corpus

For this work, we used the Pitt Corpus [25] of the DementiaBank database. This corpus contains audio recordings and manual transcriptions of participants undertaking the standard Cookie Theft picture description task from the Boston Diagnostic Aphasia Examination [14]. This password-protected data set is available upon request for research purposes.

The participants of the corpus are mainly HCs, probable and possible AD patients, and mild cognitively impaired (MCI) subjects. We excluded other diagnoses from this study due to their scarce numbers in the corpus. In this work, we did not differentiate between probable and possible diagnoses of AD. The main inclusion criterion for our study was that both the transcripts and audio files of the participant were present for each test. We studied 262 participants, with a total of 517 tests (see Table 2). Twenty-five other HC subjects and their tests were set aside for creating the referent. These subjects were not part of the experimentation sample.

2.2. Extraction of information coverage measures

2.2.1. Adaptation of the coverage measure

Velazquez's [27] measure uses duplets of linguistic patterns to find the degree to which a referent R is covered by a subject of comparison S. We selected the active voice patterns proposed by Velazquez, given the expository speech nature used during picture description tasks (see Table 3).

Velazquez splits the text into sentences; for our study, we split it into utterances. The comparison of utterance patterns follows the equation:

$$\text{coverage}(R,S) = \frac{\sum_{p\in\{R\}} MaxSim(p,S) \times \alpha_p}{\sum_{p\in\{R\}} \alpha_p}$$

where R is the referent, S is the document that is the subject of comparison, and p is a linguistic pattern. The parameters, α, are used to modify each pattern's weight. For this work,

Table 1
Linguistic characteristics selected to evaluate patients' language functions

Measure	Equation	Interpretation
Text size	N	Number of words used in a text
Vocabulary size	V	Number of different lemmas*
Hapax legomena	V_1	Number of lemmas mentioned only once
Hapax dislegomena	V_2	Number of lemmas mentioned exactly twice
Brunet's W index [35]	$W = N^{V^{-c}}$	Rationalization of the size of the vocabulary and the length of the text. W is stable when c has values between 0.165 and 0.172 [36]. We used $c = 0.172$, the original value proposed by Brunet.
Honoré's R statistics [37]	$\mathrm{R} = 100 \cdot \log N/1 - V_1/V$	A measure based on the ratio of hapax legomena, vocabulary size, and the length of the text
Type-token ratio (TTR)	$TTR = V_1/V$	TTR measures the ratio of hapax legomena and the size of the vocabulary. It can be sensitive to the size of the sample [38].
Sichel's S [39]	$S = V_2/V$	Similar to TTR, but using hapax dislegomena, being more robust against samples of different sizes [40]
Yule's characteristic K [41]	$K = 10^4 \frac{\left[\sum_{i=1}^{N} i^2 V(i,N)\right]}{N^2} - \frac{1}{N}$	Yule's is a measure of lexical repetition considered to be text length independent. In this measure, the number of lemmas of frequency i (V (i,N)) is estimated to measure the frequency distribution of a text.
Entropy	$H(X) = -\sum_{x\in X} p(x)\cdot\log_2 p(x)$	Entropy measures the uniformity of the vocabulary. In the equation, $p(x)$ is the probability of a word x occurring in the text X. We measured the general entropy of the complete text and the average entropy of sentences.

*Lemmas refer to words without inflections (in their canonical form).

Table 2
Distribution of interviews used for experimentation

Variable	All (n = 517)	AD (n = 257)	HC (n = 217)	MCI (n = 43)
Participants	262	169	74	19
Gender				
Male	189	87	75	27
Female	328	170	142	16
Education (years)				
6–9	55	51	2	2
10–12	200	112	79	9
13–16	209	76	111	22
17+	53	18	25	10
Age (years)				
Under 50	6	0	5	1
50–59	81	21	57	3
60–69	188	81	94	13
70–79	190	111	57	22
80+	52	44	4	4

Abbreviations: AD, Alzheimer's disease; HC, healthy elderly control; MCI, mild cognitive impairment; n, number of tests.

all patterns were considered to weigh equally; all parameters, α_p, were thus set to 1.

2.2.2. Automatic preprocessing

We cleaned the original raw text to apply the information coverage measure as follows:

1. We removed all marks of repetitions, hesitations, incomplete words, and pauses.
2. We standardized the names of the most prominent ICUs. For example, all mentions of the words "*brother*", "*lad*", "*kid*", etc., were automatically replaced by "*boy*" following Fig. 1.
3. We used FreeLing 4.0 [39] for tagging the transcripts with their *lemmas* and their part of speech.
4. Two consecutive nouns were considered as a single noun.
5. Some authors have found differences in the use of adjectives between AD patients and HCs [8]. During the picture description task, it is common that participants describe objects with adjectives. To take these rich descriptions into account, we joined an adjective preceded by the verb "*to be*" by means of a forward slash.
6. All part-of-speech tags that were not in the linguistic patterns were discarded for the comparison.

Table 3
Active voice linguistic patterns used for the coverage measure (reproduced from Velazquez [27])

p in *R*	*p* in *S*	Interpretation	Example
N-V	N-V	Subject + action	"boy stealing"
V-N	V-N	Action over an object	"stealing cookies"
P-N	P-N	Locations + indirect objects	"in kitchen"
N-V	V-N	Subject + action + object	"woman washing dishes"

Abbreviations: *p*, pattern; *R*, referent; *S*, subject of comparison; N, noun; V, verb; P, preposition.

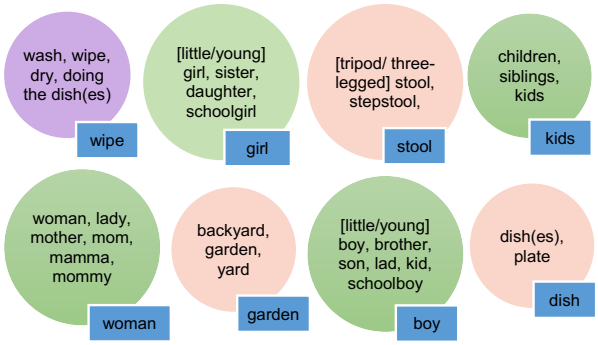

Fig. 1. Linguistic variations of ICUs in the Cookie Theft picture description task. The standardized name of each group is shown. Abbreviation: ICUs, information content units.

2.2.3. Creation of the referent

Using Velazquez's coverage measure, we created a referent that included the patterns extracted from tests taken by HCs from the same corpus. These entrances were excluded from the evaluation sample. To create the referent, we selected all 25 HCs tested only once, aiming for the referent to be as diverse as possible, while simultaneously avoiding reducing significantly the number of samples left for the evaluation.

For each utterance, if the utterance was not already at least 80% covered by the referent, the patterns were added to the referent. Thus, we automatically created an incremental referent that considered different manners used by HCs to describe similar actions and situations. The following are real examples of patterns in the referent:

water(N) *run*(V)
water(N) *overflow*(V)
water(N) *spill*(V)
water(N) *flow*(V)
water(N) *splash*(V)
spill(V) *water*(N)
kitchen/water(N) *overflow*(V)

2.2.4. Scoring participants' performance

Informativeness was estimated by measuring how much of the information in the referent was covered by the participant. To measure the *pertinence*, we estimated how much of what the participant said was covered by the referent. A low pertinence coverage may indicate that the participant was drifting off-topic.

Emulating a typical clinical scoring of a picture description test, we counted the number of utterances from the referent that exceeded an *informativeness* threshold and a *pertinence* threshold. We tested three different thresholds: 60%, 80%, and 100%. We also estimated the sum of the *informativeness* and *pertinence*.

2.3. Extraction of linguistic and phonetic characteristics

For extracting the linguistic characteristics, we conducted a usual natural-language-processing preprocessing by removing

all marks of repetitions, hesitations, incomplete words, and pauses. We used FreeLing 4.0 [39] to automatically tag the transcripts with their lemmas and part of speech. We then automatically extracted the linguistic characteristics mentioned in Section 1.2.

We used python_speech_features 0.6 [40] to estimate the first 13 MFCC values of the sound waves in 25-ms segments. As per Fraser et al. [24], our features consisted of the mean, kurtosis, skewness, and variance of the values of each MFCC.

2.4. Automatic classification

To automatically discriminate between HCs and cognitively impaired individuals, we used two widely recognized machine learning algorithms, namely, Support Vector Machine (SVM) [41] and Random Forests Classifier [42]. In Asgari et al.'s study [43], a succinct and elegantly simplified explanation of both algorithms and of their use in a linguistic analysis for detecting MCI is presented.

For our evaluation, we performed two binary classification experiments: first, a classification between participants with AD and HCs; then, we added the MCI participants to the sample and classified HCs and cognitively impaired participants. The sample of MCI participants was too small to be used as a learning class.

For this work, we used 90% of the evaluation sample as the training set and 10% as the test set. We performed a 10-fold cross-validation (see Fig. 2). We report average of the 10 test classifications.

3. Results

3.1. Feature analysis

Our evaluation of participants' picture descriptions covered a total of 105 features, divided into information coverage measures and linguistic and phonetic characteristics. We estimated their correlation with the severity of the cognitive impairment diagnosis (healthy = 0, MCI = 1, and AD = 2) and with the Mini–Mental State Examination results of participants. These correlations are reported in Table 4.

We additionally analyzed the correlation of the features with respect to age, gender, and education. Our findings are reported in Table 5.

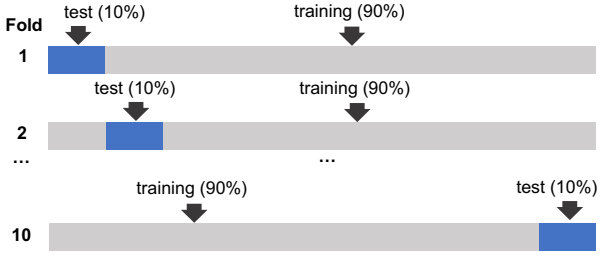

Fig. 2. Data set partitioning during a 10-fold cross-validation process to evaluate classifiers. The blue section indicates the part of the dataset that is being used as test, while the remaining gray area indicates the part of the dataset being used as training set in each fold.

3.2. Binary classification

We tested the performance of the algorithms with each type of feature independently and then combinations of them. The results of the first and second experiments are shown in Tables 6 and 7, respectively. The best model represents the performance of the algorithms with a higher area under the curve of receiver operating characteristics (AUC) during the 10-fold cross-validation process.

4. Discussion

We presented a methodology for a computer-based evaluation of a picture description task. This evaluation aims not only to score participants' performance during the task itself but also to analyze their language and phonetic productions in a single commonly used noninvasive clinical test. Our objective is to provide clinicians with computational aids for the early detection of signs that might alert of the presence of MCI or AD.

From the features observed in Table 4, the strongest correlations with the severity of the cognitive impairment were obtained with the information coverage measures. The less informative or pertinent the picture description, the higher the severity of the impairment. These correlations were consistent with the participants' scores on the Mini–Mental State Examination and were mostly independent of age, gender, and education.

Our findings on the correlation of linguistic and phonetic features with cognitive impairment were consistent with previous literature and provided a broader evaluation of the

Table 4
Correlations* of features with the severity of cognitive impairment and with the MMSE

Correlation to cognitive impairment		Correlation to MMSE score	
Variable	Corr.†	Variable	Corr.†
Informativeness t = 100%	−0.408	Informativeness t = 100%	0.443
Informativeness t = 80%	−0.388	Informativeness t = 80%	0.437
Informativeness score	−0.334	Informativeness score	0.429
Informativeness t = 60%	−0.333	Informativeness t = 60%	0.372
Informativeness variance	−0.257	Informativeness variance	0.338
Hapax legomena	−0.254	Auxiliary verb frequency	0.305
Pertinence t = 100%	−0.222	Hapax legomena	0.265
Auxiliary verb frequency	−0.216	Auxiliary verb rate	0.241
MFCC-12 kurtosis	0.205	Noun frequency	0.226
Pertinence t = 80%	−0.201	Preposition rate	0.194
MFCC-8 kurtosis	0.198	Pertinence t = 100%	0.192
MFCC-12 skewness	−0.185	General entropy	0.189
Noun frequency	−0.183	Vocabulary size	0.187
Honoré's *R* statistics	−0.180	Pertinence t = 80%	0.183
MFCC-10 kurtosis	0.163	Honoré's *R* statistics	0.177
Conjunction rate	0.163	Preposition frequency	0.175
Vocabulary size	−0.156	MFCC-12 skewness	0.173

Abbreviations: MMSE, Mini–Mental State Examination; Corr, correlation; t, threshold; MFCC, Mel-Frequency Cepstral Coefficient.

*All correlations presented with P value $< .001$. Variables are shown in descending order with respect to the strength of their correlation.

†Controlled for education, age, and gender.

Table 5
Correlations* of features with socioeconomic variables

Age		Gender		Education	
Variable	Corr.†	Variable	Corr.‡	Variable	Corr.§
MFCC-3 kurtosis	−0.200	MFCC-10 kurtosis	−0.239	Preposition freq.	0.230
Conjunction freq.	0.182	MFCC-12 variance	0.181	Hapax legomena	0.222
Brunet's *W* index	0.179	MFCC-13 variance	0.179	Vocabulary size	0.219
General entropy	0.177	MFCC-5 skewness	0.175	Text size	0.207
Auxiliary verb freq.	0.175	MFCC-5 variance	0.174	General entropy	0.201
MFCC-1 variance	0.172	MFCC-8 skewness	−0.169	Adjective freq.	0.200
MFCC-6 kurtosis	−0.168	MFCC-10 skewness	0.163	Conjunction freq.	0.191
MFCC-5 kurtosis	−0.164			Noun freq.	0.190
MFCC-9 kurtosis	−0.158			Informativeness t = 60%	0.190
Informativeness score	0.156			Auxiliary verb freq.	0.184
				Verb freq.	0.172
				Informativeness score	0.169
				Brunet's *W* index	0.167

Abbreviations: Corr, correlation; MFCC, Mel-Frequency Cepstral Coefficient; freq., frequency.
*All correlations presented with *P* value < .001. Variables are shown in descending order with respect to the strength of their correlation.
†Controlled for education, gender, and cognitive status.
‡Controlled for age, education, and cognitive status.
§Controlled for age, gender, and cognitive status.

participants' performance. In general, vocabulary richness and syntactic complexity measures were inversely correlated with the severity of the disease. These variables were also positively correlated with the number of years of education.

Phonetic variables were highly correlated with age and gender. We also observed an increase in the entropy of the description, which may indicate more chaotic or disorganized descriptions.

An increased rate of conjunctions was also positively correlated with cognitive impairment and age. The use of coordinating conjunctions in spoken language is not necessarily an indicator of grammatical connections. These coordinators in speech often have a "loose discoursal linking function" [44] and other pragmatic functions [45]. Hence, an increased use of conjunctions does not automatically imply a more complex discourse. In this work, we observed

Table 6
Performance of classifiers separating HCs from AD patients

Learner	Features	Accuracy	Sensitivity	Specificity	Precision	F-score	AUC
Average performance							
RFC	Ling	0.72	0.76	0.67	0.74	0.75	0.72
SVM	Ling	0.75	0.75	0.74	0.77	0.76	0.75
RFC	Cov	0.73	0.78	0.67	0.73	0.75	0.72
SVM	Cov	0.74	0.80	0.67	0.74	0.77	0.74
RFC	Phon	0.59	0.66	0.52	0.62	0.64	0.59
SVM	Phon	0.62	0.70	0.52	0.63	0.66	0.61
RFC	Cov + Ling	0.78	0.84	0.72	0.78	0.81	0.78
SVM	Cov + Ling	**0.79**	**0.79**	**0.78**	**0.82**	**0.80**	**0.79**
RFC	Best*	0.75	0.78	0.71	0.76	0.77	0.74
SVM	Best*	**0.79**	**0.81**	**0.77**	**0.81**	**0.81**	**0.79**
Best model							
RFC	Ling	0.81	0.77	0.86	0.87	0.82	0.82
SVM	Ling	0.85	0.85	0.86	0.88	0.86	0.85
RFC	Cov	0.85	0.88	0.82	0.85	0.87	0.85
SVM	Cov	0.85	0.88	0.82	0.85	0.87	0.85
RFC	Phon	0.67	0.65	0.68	0.71	0.68	0.67
SVM	Phon	0.72	0.84	0.57	0.70	0.76	0.71
RFC	Cov + Ling	**0.94**	**1.00**	**0.86**	**0.90**	**0.95**	**0.93**
SVM	Cov + Ling	0.88	0.81	0.95	0.95	0.88	0.88
RFC	Best*	0.85	0.85	0.86	0.88	0.86	0.85
SVM	Best*	0.87	0.80	0.95	0.95	0.87	0.88

Abbreviations: AD, Alzheimer's disease; AUC, area under the curve of receiver operating characteristics; HCs, healthy elderly controls; RFC, Random Forests Classifier; SVM, Support Vector Machine classifier; Ling, set of all linguistic features; Cov, set of all information coverage features; Phon, set of all phonetic features; Cov + Ling, a combination of all linguistic and information coverage features.
NOTE. The best results are indicated in bold.
*A combination of all features with *P* value < .001 when correlating with cognitive impairment.

Table 7
Performance of classifiers separating HCs from cognitively impaired patients (AD or MCI, indistinctly)

Learner	Features	Accuracy	Sensitivity	Specificity	Precision	F-score	AUC
Average performance							
RFC	Ling	0.70	0.78	0.59	0.73	0.75	0.69
SVM	Ling	0.72	0.80	0.61	0.74	0.77	0.70
RFC	Cov	0.74	0.83	0.61	0.75	0.79	0.72
SVM	Cov	0.73	0.86	0.56	0.73	0.79	0.71
RFC	Phon	0.59	0.79	0.31	0.61	0.69	0.55
SVM	Phon	0.61	0.81	0.33	0.62	0.70	0.57
RFC	Cov + Ling	0.76	0.84	0.66	0.77	0.81	0.75
SVM	Cov + Ling	**0.78**	**0.85**	**0.68**	**0.78**	**0.82**	**0.76**
RFC	Best*	0.77	0.82	0.69	0.78	0.80	0.75
SVM	Best*	0.75	0.82	0.65	0.76	0.79	0.73
Best model							
RFC	Ling	0.78	0.80	0.76	0.83	0.81	0.78
SVM	Ling	0.87	0.87	0.86	0.90	0.88	0.87
RFC	Cov	0.83	0.87	0.77	0.84	0.85	0.82
SVM	Cov	0.83	0.90	0.73	0.82	0.86	0.81
RFC	Phon	0.67	0.90	0.36	0.66	0.76	0.63
SVM	Phon	0.65	0.90	0.29	0.64	0.75	0.59
RFC	Cov + Ling	0.85	0.87	0.82	0.87	0.87	0.84
SVM	Cov + Ling	0.85	0.87	0.82	0.87	0.87	0.84
RFC	Best*	**0.87**	**0.87**	**0.86**	**0.90**	**0.88**	**0.87**
SVM	Best*	0.83	0.83	0.82	0.86	0.85	0.83

Abbreviations: AD, Alzheimer's disease; AUC, area under the curve of receiver operating characteristics; HCs, healthy elderly controls; RFC, Random Forests Classifier; SVM, Support Vector Machine classifier; Ling, set of all linguistic features; Cov, set of all information coverage features; Phon, set of all phonetic features; Cov + Ling, a combination of all linguistic and information coverage features; MCI, mild cognitive impairment.

NOTE. The best results are indicated in bold.

*A combination of all features with P value $< .001$ when correlating with cognitive impairment.

that a high use of conjunctions in a picture description task may even indicate hesitation. For example, in the participant's description, "she has water on the floor and… and basically, it's kind of—uh—a distressing scene," the participant appears to be repeating the conjunction "and" to gain time to further evaluate the scene.

We tested SVM and Random Forests Classifier first with each type of feature independently, and then with combinations of same. When we experimented with all three types of features together, we carried out a preselection of the best features. For this selection, we chose features with $P < .001$ when correlating to the severity of the cognitive impairment.

4.1. Comparison to other approaches

Contrasting our results against previous works on automated evaluation of picture description tasks can be difficult for multiple reasons. First and foremost, it is not customary in natural language processing to provide performance metrics such as AUC and specificity. Although accuracy, precision, recall, and F-score are usually illustrative in classes with similar sample sizes, these values could become misleading when the classes are skewed.

An additional challenge in contrasting these methods is that not every author works with the same data distribution even when using the same data set. With machine learning algorithms, the ways the samples are distributed along the data set and in the training and test sets lead to slightly different results. Authors tend to report the results obtained with a distribution in which their algorithms performed at their best.

Finally, despite using the Pitt Corpus, previous works differ in the number of samples used during their evaluation. Fraser et al. [24] used 233 HC and 240 AD samples; Yancheva and Rudzicz [26] used 241 HC and 255 AD samples; for this work, we used 242 HC and 257 AD samples (about 10% of the HC sample was used to form the referent and was not included in the evaluation). There is no clear explanation from previous authors regarding why they did not include all the samples in their experimentation.

Fraser et al. [24] reported an accuracy of 81.92%, whereas Yancheva and Rudzicz [26] reported an accuracy, precision, recall, and F-score of 80%. In both works, the authors performed a classification between HCs and AD participants, without including the MCI sample. In our work, two SVM classifiers tied with the highest AUC at 0.79 in this task (Table 6). The first learner used a combination of all the information coverage and linguistic features, whereas the second used a combination of all features with $P < .001$ when correlating with cognitive impairment. The second algorithm presented a higher sensitivity (81%) and a higher F-score (81%), comparable to state-of-the-art work [24] that uses a manually made list of ICUs.

When we incorporated the MCI sample into the experiment (Table 7), we observed that the SVM learner, trained with information coverage and linguistic features,

performed at the highest AUC (0.76). There was an expected increase in the false-negative rate (specificity = 68%). However, the sensitivity was still high at 85%.

The best model of an experiment represents the highest performance achieved by an algorithm during the cross-validation process. This indicates the highest potential of the algorithms in classifying new data with similar characteristics to the sample. For the first experiment (Table 6), the best model had an AUC of 0.93, with excellent sensitivity and a true-negative rate of 86% when classifying HCs and AD patients. When the MCI sample was incorporated (Table 7), the best model had an AUC of 0.87, with sensitivity 87% and specificity 86%.

4.2. Study advantages and limitations

One of the advantages of our proposed methodology is that the informativeness and pertinence measures are estimated against an automatically created referent. This referent has the particularity of being adaptable to differences in population or even to different pictures for description.

Previous automated works present difficulties at considering linguistic variabilities for expressing similar notions. With our proposed approach, the referent is created from examples of descriptions from healthy age-related individuals. Hence, it incorporates different ways of expressing similar ideas and even what could be considered as *normal* deviations from topics. It also allows for the consideration of context through the accounting of linguistic patterns of phrases, rather than just of isolated words. In this regard, the bigger the sample set aside for creating the referent, the richer and more variate the referent.

To our knowledge, this is the first time that an automatic measure of pertinence has been implemented in a picture description task. While most computational approaches focus only on the information coverage, one advantage of our measure is that it helps to detect when patients drift off-topic, a highly challenging task in automatic analysis.

One disadvantage of our approach is that it sacrifices part of the HC group to create the referent, reducing the availability of HC samples for training the algorithms. Our study also presented a limitation when evaluating MCI patients, yielding a high false-negative rate.

4.3. Future work

In both experiments, we observed that our phonetic characteristics were not sufficiently discriminative or had little to no effect in the performance of the algorithms. As previous authors have reported, the use of more complex acoustic and rhythm features could significantly increase the automatic classification performance of HCs and AD patients.

In future work, we propose to extend the research scope with the evaluation of the performance of the information coverage metrics in descriptions of different picture description tasks or even in different restricted-discourse tests. Also, there is a potential to perform multilingual studies because all the features proposed in this work are language independent or can be adapted for studies in different languages. Finally, we intend to research the effects of different HC sample sizes for creating the referent.

Acknowledgments

This research was partially funded by the scholarships FRQNT-177601 and CONACYT-231979, as well as the joint project of the *Ministère des Relations internationales et de la Francophonie* Quebec-CONACYT, and the PAPIIT project IA400117.

RESEARCH IN CONTEXT

1. Systematic review: Alzheimer's disease (AD) patients and healthy elderly controls have shown significant differences in their performance on picture description tests. We present a computer-based methodology to evaluate the performance of patients during this task.
2. Interpretation: Using 10% of the healthy elderly controls and an information coverage measure, we created a population-specific referent. Against this referent, we automatically assessed the informativeness and pertinence of descriptions of the Cookie Theft picture and extracted linguistic and phonetic features. Applying machine learning algorithms, we classified healthy elderly controls and AD patients with results (sensitivity = 81%; AUC = 0.79) comparable to state-of-the-art work that uses a manually made referent. We incorporated mild cognitively impaired participants (sensitivity = 85%; AUC = 0.76). Our findings encourage the use of computer-based procedures as an aid in clinical practice.
3. Future directions: Our proposed approach could be applied to other picture tasks and languages. Additional studies are needed with larger mild cognitively impaired samples and more complex phonetic features to improve the classification accuracy.

References

[1] Szatloczki G, Hoffmann I, Vincze V, Kalman J, Pakaski M. Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. Front Aging Neurosci 2015;7:195.
[2] Kaplan E, Goodglass H, Weintraub S. Boston Naming Test. Philadelphia, PA: Lea & Febiger; 1983.

[3] De Renzi E, Vignolo LA. The token test: A sensitive test to detect receptive disturbances in aphasics. Brain 1962;85:665–78.
[4] Strauss E, Sherman EMS, Spreen O. A compendium of neuropsychological tests: Administration, norms, and commentary. 3rd ed. Victoria, BC, Canada: Oxford University Press, Department of Psychology, University of Victoria; 1998.
[5] Laske C, Sohrabi HR, Frost SM, López-de-Ipiña K, Garrard P, Buscema M, et al. Innovative diagnostic tools for early detection of Alzheimer's disease. Alzheimers Dement 2015;11:561–78.
[6] Bucks RS, Singh S, Cuerden JM, Wilcock GK. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. Aphasiology 2000;14:71–91.
[7] Alegria R, Gallo C, Bolso M, dos Santos B, Prisco CR, Bottino C, et al. Comparative study of the uses of grammatical categories: Adjectives, adverbs, pronouns, interjections, conjunctions and prepositions in patients with Alzheimer's disease. Alzheimers Dement 2013;9:P882.
[8] Jarrold W, Peintner B, Wilkins D, Vergryi D, Richey C, Gorno-Tempini ML, et al. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. Proc. ACL Work. Comput. Linguist. Clin. Psychol; 2014. p. 27–36.
[9] Khodabakhsh A, Kusxuoglu S, Demiroglu C. Natural language features for detection of Alzheimer's disease in conversational speech. IEEE-EMBS International Conference on Biomedical and Health Informatics; 2014. p. 581–4.
[10] López-de-Ipiña K, Solé-Casals J, Eguiraun H, Alonso JB, Travieso CM, Ezeiza A, et al. Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach. Comput Speech Lang 2015;30:43–60.
[11] König A, Satt A, Sorin A, Hoory R, Toledo-Ronen O, Derreumaux A, et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimer's Dement Diagnosis. Assess Dis Monit 2015;1:112–24.
[12] Fraser K, Hirst G. Detecting semantic changes in Alzheimer's disease with vector space models. Proc Lr 2016 Work Resour Process Linguist Extra-Linguistic Data from People with Var Forms Cogn Impair; 2016. p. 1–8.
[13] Zhou L, Fraser K, Rudzicz F. Speech recognition in Alzheimer's disease and in its assessment. INTERSPEECH; 2016. p. 1948–52.
[14] Goodglass H, Kaplan E. The Assessment of Aphasia and Related Disorders. Philadelphia: Lea & Febiger; 1983.
[15] Ahmed S, de Jager CA, Haigh A-M, Garrard P. Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease. Neuropsychology 2013;27:79–85.
[16] Yorkston KM, Beukelman DR. An analysis of connected speech samples of aphasic and normal speakers. J Speech Hear Disord 1980;45:27–36.
[17] Hier DB, Hagenlocker K, Shindler AG. Language disintegration in dementia: Effects of etiology and severity. Brain Lang 1985;25:117–33.
[18] Nicholas M, Obler LK, Albert ML, Helm-Estabrooks N. Empty speech in Alzheimer's disease and fluent aphasia. J Speech Hear Res 1985; 28:405–10.
[19] Croisile B, Ska B, Brabant M-J, Duchene A, Lepage Y, Aimard G, et al. Comparative study of oral and written picture description in patients with Alzheimer's disease. Brain Lang 1996;53:1–19.
[20] Forbes-McKay KE, Venneri A. Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. Neurol Sci 2005;26:243–54.
[21] Lai YH, Pai HH, Lin YT. To be semantically-impaired or to be syntactically-impaired: Linguistic patterns in Chinese-speaking persons with or without dementia. J Neurolinguist 2009;22:465–75.
[22] Hakkani-Tür D, Vergyri D, Tur G. Speech-based automated cognitive status assessment. INTERSPEECH-2010; 2010. p. 258–61.
[23] Pakhomov SVS, Smith GE, Chacon D, Feliciano Y, Graff-Radford N, Caselli R, et al. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. Cogn Behav Neurol 2010;23:165–77.
[24] Fraser K, Meltzer J, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. J Alzheimers Dis 2015;49:407–22.
[25] Becker JT, Boiler F, Lopez OL, Saxton J, McGonigle KL. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. Arch Neurol 1994;51:585–94.
[26] Yancheva M, Rudzicz F. Vector-space topic models for detecting Alzheimer's disease. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany; 2016. p. 2337–46.
[27] Velázquez-Godínez E. Caractérisation de la couverture d'information: Une approche computationnelle fondée sur les asymétries. Montreal, Canada: École de technologie supérieure, Quebec University; 2017.
[28] Snowdon DA, Kemper SJ, Mortimer JA, Greiner LH, Wekstein DR, Markesbery WR. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Finds from the Nun Study. J Am Med Assoc 1996;275:528–32.
[29] Jarrold WL, Peintner B, Yeh E, Krasnow R, Javitz HS, Swan GE. Language Analytics for Assessing Brain Health: Cognitive Impairment, Depression and Pre-symptomatic Alzheimer's Disease. Toronto, Canada: Lecture Notes in Computer Science (Including its subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); 2010. p. 299–307.
[30] Guinn C, Habash A. Language Analysis of Speakers with Dementia of the Alzheimer's Type. Association for the Advancement of Artificial Intelligence Fall Symposium; 2012. p. 8–13.
[31] Alegria R, Bolso M, Gallo C, Prisco CR, Bottino C, Ines NM. Retained lexis in people with Alzheimer's disease. Alzheimers Dement 2013; 9:P486–7.
[32] Khodabakhsh A, Yesil F, Guner E, Demiroglu C. Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech. EURASIP J Audio, Speech, Music Process 2015; 2015. p. 9.
[33] Ahmed S, Haigh A-MF, De Jager CA, Garrard P. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. Brain 2013;136:3727–37.
[34] Kemper S, LaBarge E, Ferraro R, Cheung H, Cheung H, Storandt M. On the preservation of syntax in Alzheimer's disease: Evidence from written sentences. Arch Neurol 1993;50:81–6.
[35] Rudzicz F, Chan Currie L, Danks A, Mehta T, Zhao S. Automatically Identifying Trouble-indicating Speech Behaviors in Alzheimer's Disease. Proceedings of the 16th international ACM SIGACCESS Conference on Computers & Accessibility; 2014. p. 241–2.
[36] Satt A, Hoory R, König A, Aalten P, Robert PH, Sophia N, et al. Speech - Based Automatic and Robust Detection of Very Early Dementia. INTERSPEECH; 2014. p. 2538–42.
[37] Lopez-de-Ipina K, Martinez-de-Lizarduy U, Barroso N, Ecay-Torres M, Martinez-Lage P, Torres F, et al. Automatic analysis of Categorical Verbal Fluency for Mild Cognitive impartment detection: A non-linear language independent approach. 4th IEEE International Work-Conference on Bioinspired Intelligence; 2015. p. 101–4.
[38] Khodabakhsh A, Demiroglu C. Analysis of speech-based measures for detecting and monitoring Alzheimer's disease. Methods Mol Biol 2015;1246:159–73.
[39] Padró L, Stanilovsky E. Freeling 3.0: Towards wider multilinguality. Association ELR. Istanbul, Turkey: Proceedings of the Language Resources and Evaluation Conference; 2012. p. 2473–9.
[40] Lyons J. python_speech_features 0.6; 2017. Available at: https://pypi.python.org/pypi/python_speech_features. Accessed March 15, 2018.
[41] Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput 2004;14:199–222.
[42] Breiman L. Random Forests. Mach Learn 2001;45:5–32.
[43] Asgari M, Kaye J, Dodge H. Predicting mild cognitive impairment from spontaneous spoken utterances. Alzheimer's Dement Transl Res Clin Interv 2017;3:219–28.
[44] Leech G. Grammars of spoken english: New outcomes of corpus-oriented research. Lang Learn 2000;50:675–724.
[45] Jørgensen F. Clause Boundary Detection in Transcribed Spoken Language. Proceedings of the 16th Nordic Conference of Computational Linguistics; 2007. p. 235–9.