Linguistic Parameters of Spontaneous Speech for Identifying Mild Cognitive Impairment and Alzheimer Disease

Veronika Vincze
MTA-SZTE Research Group on
Artificial Intelligence
vinczev@inf.u-szeged.hu

Martina Katalin Szabó
MTA TK Computational Social Science -
Research Center for Educational and
Network Studies (CSS-RECENS)
and University of Szeged
Institute of Informatics
martina@inf.u-szeged.hu

Ildikó Hoffmann
Research Centre for Linguistics
Eötvös Lorand Research Network
and University of Szeged
Department of Hungarian
Linguistics, Szeged
hoffmannildi@gmail.com

László Tóth
University of Szeged
Institute of Informatics
tothl@inf.u-szeged.hu

Magdolna Pákáski
University of Szeged
Department of Psychiatry
babikne.pakaski.magdolna
@med.u-szeged.hu

Submission received: 9 July 2020; revised version received: 14 September 2021; accepted for publication: 5 December 2021.

https://doi.org/10.1162/COLI_a_00428

János Kálmán
University of Szeged
Department of Psychiatry
kalman.janos@med.u-szeged.hu

Gábor Gosztolya
MTA-SZTE Research Group on
Artificial Intelligence
ggabor@inf.u-szeged.hu

In this article, we seek to automatically identify Hungarian patients suffering from mild cognitive impairment (MCI) or mild Alzheimer disease (mAD) based on their speech transcripts, focusing only on linguistic features. In addition to the features examined in our earlier study, we introduce syntactic, semantic, and pragmatic features of spontaneous speech that might affect the detection of dementia. In order to ascertain the most useful features for distinguishing healthy controls, MCI patients, and mAD patients, we carry out a statistical analysis of the data and investigate the significance level of the extracted features among various speaker group pairs and for various speaking tasks. In the second part of the article, we use this rich feature set as a basis for an effective discrimination among the three speaker groups. In our machine learning experiments, we analyze the efficacy of each feature group separately. Our model that uses all the features achieves competitive scores, either with or without demographic information (3-class accuracy values: 68%–70%, 2-class accuracy values: 77.3%–80%). We also analyze how different data recording scenarios affect linguistic features and how they can be productively used when distinguishing MCI patients from healthy controls.

1. Introduction

Alzheimer disease (AD) is a neurodegenerative disorder that develops for years before clinical manifestation, while mild cognitive impairment (MCI) is usually viewed as a prodromal stage of AD (Galvin and Sadowsky 2012). Symptoms such as language dysfunctions may even occur nine years before the actual diagnosis (APA 2000). Thus, the language use of the patient may suggest MCI well before the clinical diagnosis of dementia. For both types of neurodegenerative disorders, an early diagnosis is crucial in order to allow timely treatment to decelerate progression (Nelson and Tabet 2015). However, according to Boise at al., for many MCI patients (up to 50%) MCI is never recognized (Boise, Neal, and Kaye 2004). A reason for this might be that in the early stages of the disease it is not easy for experts to detect cognitive impairment.

Tests that are the most sensitive to cognitive and linguistic changes occurring in early AD and other types of dementia have been intensively studied (Chapman et al. 2002). Several screening tests aim for the early detection of dementia, but they are either too time-consuming or cannot diagnose preclinical stages. For instance, diagnostic tools such as volumetric MRI (Scheltens et al. 2002; Zimny et al. 2011; Yin et al. 2013) and diffusion tensor imaging (Nakata et al. 2009; Stricker et al. 2009; Matsuda, Asada, and Tokumaru 2017) may be effective, but these are time-consuming and costly techniques for early screening. Most dementia filter tests (Mini-Mental State Examination [MMSE], Clock Drawing Test [CDT], Alzheimer's Disease Assessment Scale-cognitive subscale

[ADAS-cog]) are not able to accurately recognize MCI (Folstein, Folstein, and McHugh 1975; Rosen, Mohs, and Davis 1984; Janka et al. 1988; Kálmán, Maglóczky, and Janka 1995; Patocskai et al. 2014). Tests on linguistic memory prove more effective in detecting MCI, but they tend to yield a relatively high number of false-positive diagnoses (Roark et al. 2011). Hence, cheap but still effective methods for identifying dementia as early as possible are urgently required.

Conversation analysis has proven to be an encouraging method in detecting memory complaints (Mirheidari et al. 2017, 2016). MCI is known to affect the speech of the patient via three main aspects. First, verbal fluency declines, which results in longer hesitations and a lower speech rate (Roark et al. 2011; Pistono et al. 2019). Second, the lexical frequency of words and the differences in the frequencies of parts of speech may also change significantly as the patient has difficulties with finding lexical items (Croot et al. 2000). Third, the emotional responsiveness of the patient has also been reported to change frequently (López-de-Ipiña et al. 2015).

In connection with the above-mentioned features, researchers recently experimented with detecting different types of dementia using Automatic Speech Recognition (ASR) tools in several studies. Just to name a few, ASR tools were utilized to detect MCI (Lehr et al. 2012) and AD (Baldas et al. 2010; López-de-Ipiña et al. 2013; Satt et al. 2014; López-de-Ipiña et al. 2015; Al-Hameed et al. 2017; König et al. 2015; Weiner, Herff, and Schultz 2016). Jarrold et al. relied on speech rate and mean and standard deviation of vowels and consonants in spontaneous speech samples (Jarrold et al. 2014). Al-Hameed et al. (2017) sought to predict a common clinical examination score for dementia using acoustic information extracted from people describing a picture. They also sought to develop a diagnostic tool that is able to distinguish sufferers with AD from those with MCI and healthy controls. Their classification model is capable of predicting dementia with an average cross-visit accuracy ranging from 89.2% to 92.4% when performing pairwise classification among the AD, MCI, and healthy control classes. Al-Hameed et al. (2019) examined 15 patients with progressive neurodegenerative disorders and 15 with functional memory disorder and, based on 51 acoustic features extracted from the recordings, they identified the most discriminating features. Then these features were used to train five different machine learning classifiers to differentiate between the two classes, which gave a mean classification accuracy of 96.2%.

Types of speech production tasks have also been investigated from the viewpoint of the prediction of lexical and semantic impairment. Pistono et al. (2019) compared pause duration and frequency in the AD participants and healthy controls using a picture-based narrative and memory-based narrative. The results indicated that participants with AD had more pauses only in the picture-based narrative.

As for natural language processing (NLP) methods, the lexical analysis of spontaneous speech may also suggest different types of dementia (Holmes and Singh 1996; Bucks et al. 2000; Lunsford and Heeman 2015) and the results of these analyses can be exploited in the automatic detection of patients suffering from dementia (Thomas et al. 2005; Jarrold et al. 2014; Shibata, Wakamiya, and Aramak 2016; König et al. 2015). Changes in the writing style of people may also indicate dementia (Garrard et al. 2005; Hirst and Wei Feng 2012; Le et al. 2011). Fraser et al. were able to distinguish MCI speakers from healthy older adults with accuracy scores of up to 63% (English) and 72% (Swedish) on the basis of information content alone (Fraser, Fors, and Kokkinakis 2018). The results of these studies are very encouraging. For instance, Fraser, Fors, and Kokkinakis (2018) established that subtle differences in language can be detected in narrative speech, even at the very early stages of cognitive decline, when scores on screening tools such as the MMSE are still in the "normal" range.

Besides English, there are studies that seek to identify dementia in native speakers of, for example, German (Weiner, Herff, and Schultz 2016), Portuguese (dos Santos et al. 2017), Japanese (Shibata, Wakamiya, and Aramak 2016), and Swedish (Kokkinakis et al. 2017; Fraser et al. 2017). Fraser, Fors, and Kokkinakis (2018) analyzed the information content of narrative speech samples from individuals with MCI, in both English and Swedish, using a combination of supervised and unsupervised learning techniques. They found that the multilingual approach leads to significantly better classification accuracy scores than training on the target language alone. As for the automatic detection of MCI in Hungarian individuals, Vincze et al. (2016) sought to identify MCI patients based on linguistic features gained from the transcripts of spontaneous speech recordings. As regards speech features, Tóth et al. (2015) and Tóth et al. (2018) experimented with speech recognition techniques. To extend the previous studies concerning the Hungarian language, Gosztolya et al. (2019) involved both mild AD (mAD) and MCI patients, and speech-based and linguistic features were used in distinguishing the two classes from healthy controls.

In this article, we again seek to automatically identify Hungarian patients suffering from MCI or mAD based on their speech transcripts. In contrast with previous work (e.g., Tóth et al. 2018), here we focus on only linguistic features and ignore those derived from ASR. Our system applies machine learning techniques and is based on a rich feature set that includes parameters of linguistic characteristics of spontaneous speech along with features that exploit morphological and syntactic parsing and features derived from semantic and pragmatic phenomena. In addition to the features used in our earlier studies (Vincze et al. 2016; Gosztolya et al. 2019), we have included new morphological, syntactic, semantic, and pragmatic features that might be characteristic of spontaneous speech. We also attempt to investigate how the different data recording scenarios affect linguistic features. This also leads us to propose a methodology to identify dementia on the basis of linguistic parameters of spontaneous speech. Hence, the main contributions of the article are the following:

- We define a rich feature set of linguistic parameters for detecting different types of dementia and propose some novel features for the task;
- We carry out a detailed statistical analysis of (novel) linguistic parameters that may distinguish healthy controls (HC) from MCI and mAD patients in three different tasks, namely, immediate recall, delayed recall, and describing what happened on the previous day;
- We perform machine learning experiments with the above-mentioned feature set for detecting different types of dementia;
- We analyze the efficacy of the above-mentioned three different tasks based on the results of a data analysis from transcripts and the results of the experiments.

The article is structured as follows. In Section 2 we present the basic attributes and statistical data of the Hungarian MCI-mAD database. Then, in Section 3 we discuss the methodology of the research, along with the rich feature set applied in the processing of the speech transcripts and investigate the significance level of these values among various speaker group pairs (HC vs. MCI, HC vs. mAD, and MCI vs. mAD) for the different speaker tasks. In Section 4, we describe our machine learning experiments using the same feature set. Afterwards, in Section 5 we systematically analyze the datasets

and we show that these attributes also serve as a basis for an efffective discrimination among the three speaker groups. We will also draw some conclusions on the usefulness of each speaker task. Lastly, we summarize the main results of our study in Section 6.

2. The Hungarian MCI-mAD Database

In our study, we used the Hungarian MCI–mAD database, recorded at the Memory Clinic at the Department of Psychiatry of the University of Szeged, Hungary (Gosztolya et al. 2019). The study was approved by the Ethics Committee of the University of Szeged, and it was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all the participants involved in the research project. Unfortunately, our ethics agreement does not allow the sharing of these speech recordings. For the sake of simplicity, we will provide the most important steps of the data collection based on Gosztolya et al. (2019).

We collected utterances from three groups of subjects. Namely, those suffering from MCI, those affected by early-stage AD, and HC (i.e., those with no cognitive impairment at the time of recording). The three groups were then matched for age, gender, and education. MCI and mAD patients were selected after a medical diagnosis was confirmed by computed tomography, magnetic resonance imaging (MRI), and cognitive tests (MMSE [Folstein, Folstein, and McHugh 1975], CDT [Freedman et al. 1994], and ADAS-Cog [Rosen, Mohs, and Davis 1984]). Anyone who had previously suffered from head injuries, depression, or psychosis was excluded here. Further exclusion criteria were drug or alcohol consumption, being under pharmacological treatment affecting cognitive functions, and visual or auditory deficits. This choice is justified by the fact that head injuries may also lead to speech impairment (e.g., aphasia). Moreover, depression, alcohol use, and drug use are clinically known to affect cognitive processes, hence may influence speech as well.

Here our aim was to investigate whether we can determine the state of the patients based on linguistic features only. For this reason, we needed ground truth labels, that is, a clinically confirmed medical diagnosis for each patient, obtained in the most precise way (applying imaging processes, cognitive tests, etc.). The classification of MCI and mAD patients was always the result of a consensus between the members of our clinical expert panel (a psychiatrist, a neurologist, and a psychologist), who made their decision based on the global clinical picture, neuropsychological test results, and also neuroimaging (when available). As far as we know there is no clinical protocol for diagnosing patients only on the basis of their linguistic utterances, hence we were not able to rely on such protocols in the diagnosing phase. However, as Petersen (2004) also remarks, the distinction between healthy aging and MCI, and also between MCI and very early AD, is challenging as these conditions often overlap on a cognitive continuum. If the expert panel could not agree on the classification of a patient, that patient was not included in analysis to prevent the confounding effect of an already controversial diagnosis.

All our previous studies (Hoffmann et al. 2010; Tóth et al. 2015; Gosztolya et al. 2016; Tóth et al. 2018; Gosztolya et al. 2019) and studies carried out by other groups (e.g., Taler and Phillips 2008; Roark et al. 2011; Satt et al. 2014) found that MCI and AD affect the *spontaneous* speech of the patients more than their planned speech. In the case of planned speech, speakers usually have some time in advance to think about what they would like to say, hence difficulties in word finding (due to memory decline) cannot be reliably detected. However, in the case of spontaneous speech, speakers are required to speak on the spot, so they do not have time to prepare their speech, which might truly reflect their

Table 1
The instructions to the patients when recording the three utterances.

(1)	"*I am going to show you a silent movie lasting about a minute. Try to remember the story, the actors, the objects and the places, paying attention to the details.*"
(2)	"*Now, I would like to ask you to tell me about yesterday in detail.*
(3)	"*Now, I am going to show you another clip. Try to remember the story, the actors, the objects and the places, paying attention to the details. OK, I am going to start it now.*" The Patient watches the clip. If he starts talking about it, he is reminded that he is not yet allowed to talk about it. When the clip ends: "*Now we will take a one-minute break.*" If the Patient starts talking during the break, he is reminded that it is still break time, and he has to wait until the minute is over. After the one-minute break is over: "*Right, could you please tell me what you saw in the clip?*"

difficulties in word finding. Therefore, our aim was to record spontaneous speech, and use the transcripts of these utterances. This is why our experimental setup for recording was as follows (for the details, see Hoffmann et al. 2010). After the presentation of a specially designed one-minute-long animated film, the subjects were asked to talk about the events seen on the film (*immediate recall* or *Task 1*). Next, the subjects were asked to talk about their previous day (*previous day* or *Task 2*). As the last task, the subjects were shown a second film, then—after a one-minute long pause—were asked to talk about the second film (*delayed recall* or *Task 3*). (For the instructions to the subjects, see Table 1.) Hence, we had three recordings for each subject, each containing spontaneous speech, but the tasks performed were different. In this article, we also seek to investigate whether some tasks are less effective for detecting MCI or mAD than other tasks. This is why we experimented with three different recordings.

Our approach makes use of textual input, that is, the transcripts of utterances made by the speaker groups. However, it must be emphasized that this method may be complementary to using speech recordings as we did in our previous work (Tóth et al. 2015; Tóth et al. 2018; Gosztolya et al. 2019). We think that the combination of these two methodologies, namely, relying on textual information as well as on automatic speech recognition techniques, can lead to even higher accuracy with regard to identifying the patients' status, which we would like to implement in the future.

Our database of MCI and AD patients is continuously growing; at the time of writing we had recordings taken from more than 150 persons. For various reasons (poor sound quality, controversial diagnosis, etc.) we had to filter out some patients; furthermore, because we insisted on matching the three groups of speakers by age, gender, and level of education, we could not use some of the recordings, which otherwise fulfilled our requirements of having a clear diagnosis and an acceptable sound quality. Therefore, in the end we used the recordings of 25 speakers for each speaker group, resulting in a total of 75 speakers and 225 recordings. We applied one-way ANOVA to check if there were significant differences among the different groups. F and p-values can be seen in Table 2. It can be seen that the differences in the age and years of education are statistically not significant (p-values of 0.105 and 0.118), while the MMSE, CDT, and Adas-COG tests indeed show a statistically significant difference among the speaker groups. With t-tests, we also checked whether there are significant differences among

Table 2
Demographic data and the results of the MMSE, CDT, and Adas-Cog tests of the three subject groups. We also report mean and standard deviation (mean ± SD).

	Subject groups			Statistics	
	Control (n = 25)	**MCI** (n = 25)	**mAD** (n = 25)	F (2;74)	p
Age	70.72 ± 5.004	72.4 ± 3.594	73.96 ± 6.846	2.321	$p = 0.105$
Education	12.08 ± 2.326	10.84 ± 2.304	10.76 ± 2.818	2.202	$p = 0.118$
MMSE	29.24 ± 0.523	27.16 ± 0.898	23.92 ± 2.488	76.213	$p < 0.001$
CDT	8.88 ± 2.007	6.44 ± 3.429	5.88 ± 3.244	7.254	$p = 0.001$
Adas-COG	8.575 ± 2.374	12.044 ± 3.205	18.675 ± 5.818	38.35	$p < 0.001$

Table 3
Significance of demographic data and the MMSE, CDT, and Adas-Cog tests of healthy controls and patients with dementia.

Patient groups	**Age**	**Education**	**MMSE**	**CDT**	**Adas-COG**
Control vs. MCI	$p = 0.0912$	$p = 0.0115$	$p < 0.0001$	$p = 0.0037$	$p = 0.0002$
Control vs. MCI+mAD	$p = 0.0202$	$p = 0.0083$	$p < 0.0001$	$p < 0.0001$	$p < 0.0001$

healthy controls and patients with MCI on the one hand and healthy controls and patients with dementia (i.e., grouping the MCI and mAD patients together) on the other hand. As shown in Table 3, there are significant differences among the groups except for age in the case of the control vs. MCI speakers.

3. Methodology

In this section, we will describe our methods used to identify MCI and mAD patients based on their speech transcripts.

3.1 Feature Set

In our experiments, we used a rich feature set derived from the transcripts and the results of the automatic linguistic analyses performed with `magyarlanc`, a linguistic preprocessing toolkit for Hungarian (Zsibrita, Vincze, and Farkas 2013). With this tool, the text was first split into sentences, then tokenized, and finally the tokens were lemmatized. A token is a semantic unit, usually separated by spaces from other character sequences in the text (Szabó et al. 2020). A token can be a word, a number, or punctuation as well. Lemmatization is especially important in case of morphologically rich languages such as Hungarian. In these languages words—nouns, verbs, pronouns, and adjectives—may have numerous inflected and derived forms (Mladenović et al. 2016). This property may make the automatic analysis significantly more difficult or even ineffective. Lemmatization removes inflectional endings and returns the base or dictionary form of a word (Balakrishnan and Lloyd-Yemoh 2014; Kutuzov

and Kuzmenko 2019). As a last step of preprocessing, punctuation was removed. The remaining strings are referred here as *words*.

Similarly to Tóth et al. (2015), we hypothesized that the speech of MCI patients may contain more pauses and hesitations than the speech of HC and they are also supposed to have a restricted vocabulary due to cognitive deficit, which may affect the choice of words and the frequency of parts of speech (Croot et al. 2000), and they might even produce neologisms. In addition to the features used in our earlier study, we added new morphological, syntactic, semantic, and pragmatic features that might be characteristic of spontaneous speech, and we made use of demographic features that were available to us. Altogether, the feature set consisted of 330 features (3 demographic features and 3 times 109 features for each recording).

Our feature set contained the following features (novel features that have not been applied in our previous studies are italicized):

We extracted basic **statistical features** (7 features) from each transcript, namely:

- The number of sentences;
- The number and relative frequency of tokens;
- The number of words;
- The number and frequency of distinct lemmas compared to the number of words;
- *The average sentence length*.

We also processed the data from the viewpoint of **spontaneous speech-based features** (6 features):

- The number of filled and silent pauses;
- The number and frequency of hesitations compared to the number of tokens;
- The number of pauses that follow an article and precede content words, as this might indicate that MCI patients may have difficulties in finding the suitable content words;
- The number of lengthened sounds (which we treated as a special form of hesitation based on Gosztolya et al. [2016]).

Most of the **morphological features** employed in our analysis rely on the fact that Hungarian is a morphologically rich language, and this is why many grammatical relations are expressed by suffixes, the number of which might indicate whether or not the cognitive abilities of the speaker have been adversely affected. In this phase of the data processing we extracted the following features (35 features altogether):

Part-of-speech (or *POS*) *features* (17 features):

- The number and frequency of nouns, verbs, adjectives, pronouns, *numerals*, *adverbs*, and conjunctions compared to the number of all words;
- The number of punctuation marks;

- The number and frequency of unanalyzed words, that is, those with an "unknown" POS tag, compared to the number of all words, which could reflect whether neologisms are being created by the speaker while speaking.

Deep morphological features (18 features):

- *The number of first person singular verbs*, as this might tell us how often the patient reflects upon himself or herself;
- *The number of first person plural verbs*, as this might provide evidence for a strong or weak group identity of the patient;
- *The number and frequency of past and present tense verbs* compared to the number of all verbs, as this might reflect how well the patient can remember past events;
- *The number and frequency of imperative and conditional verbs* compared to the number of all verbs, as this might provide evidence how the patient is able to cognitively perceive non-factual events;
- *The number and frequency of comparative and superlative adjectives* compared to the number of all adjectives, as this might tell us how the patient can make comparisons;
- *The number and frequency of demonstrative pronouns* compared to the number of all pronouns, as this might indicate the ability of changing relative directions and viewpoints;
- *The average number of morphemes of nouns*.

As for **syntactic features**, we extracted the following characteristics (10 features):

- *The number and frequency of subjects and objects*, compared to the number of all words, as Hungarian is a pro-drop language, meaning that pronominal subjects and objects might not be overt in the clause;
- *The number and frequency of adverbs*, compared to the number of all words, as adverbs usually describe additional circumstances to the events and this might indicate the way the speaker recalls the story (i.e., describing only the main events or adding some further details);
- *The number and frequency of coordinations and subordinations*, compared to the number of all words, as these features may characterize the complexity of the speaker's sentences.

We also carried out an analysis of the **semantic features** of the texts from the point of view of sentiments, emotions, and words or phrases denoting uncertainty of the speaker in the veracity of the information expressed and different kinds of memory activity, among others (47 features altogether):

Uncertainty features (16 features):

- The number and frequency of fillers and uncertain words compared to the total number of tokens;

- *The number and frequency of words belonging to several classes of linguistic uncertainty* based on Vincze (2014), compared to the number of all words.

Sentiment features (10 features):

- *The number and frequency of positive and negative words* based on a list of sentiment phrases, compared to the number of all words. We applied two different Hungarian dictionaries for sentiment analysis: One list was a translation of Liu (2012), while the other one contained Hungarian slang words (Szabó 2015) (in the tables "positive/negative" and "slangPositive/slangNegative," respectively.)

Emotion features (16 features):

- *The number and frequency of words belonging to the emotions* described in Szabó, Vincze, and Morvay (2016), compared to the number of all words.

Other semantic features (5 features):

- The number and frequency of words/phrases related to memory activity (e.g., *nem emlékszem* not remember-1SG "I can't remember"), compared to the number of all words, as they directly signal problems related to memory and recall;
- The number of negation words;
- The ratio of content words and function words.

As regards **pragmatic features** of the transcripts, we processed speech act verbs and discourse markers (4 features):

- *The number and frequency of speech act verbs*, based on a manually constructed list, compared to the number of all verbs;
- *The number and frequency of discourse markers*, compared to the number of all words. To find discourse markers in the texts we applied a word list based on Dér and Markó (2007).

Lastly, we also took into consideration the **demographic features** of the speakers (3 features):

- Gender;
- Age;
- Education.

All the lists we have used in the investigation of semantic and pragmatic features are available at `https://github.com/vinczev/hungarian_lists`.

3.2 Statistical Analysis of Features

In order to quantify the usefulness of each feature in distinguishing HC, MCI patients, and mAD patients, we carried out a statistical analysis of the data (pairwise t-tests for each feature and transcript). The significance levels for each feature among the three groups are listed in Tables 7 and 8, and the significance levels for each feature between HC and speakers with either MCI or mAD are listed in Tables 10 and 11, according to the following notation:

- *: $0.01 \leq p < 0.05$,
- **: $0.001 \leq p < 0.01$, and
- ***: $p < 0.001$.

The features that do not exhibit significant differences have been omitted from the tables for the sake of simplicity.

Analyzing the single features, Tables 10 and 11 tell us that almost all the features display significant differences when working with only two classes: There are only 9 features—out of 109—that do not exhibit significant differences in any of the three tasks. Hence, the use of linguistic features for distinguishing between HC and speakers with MCI or mAD is well justified and our feature set for the machine learning experiments will be based on them (see Section 4).

Figure 1 shows the results of the analysis, in accordance with the task types. More precisely, we can see how many features of the specific feature group exhibit significant differences with $p < 0.05$ for each speaker group pairs.

From the statistically significant features, we can conclude that Task 2, namely, the description of the previous day, proves to be the best indicator to differentiate between HC and speakers with MCI. On the other hand, Task 3 is useful when patients with MCI and mAD need to be distinguished. A more detailed analysis of feature groups and the effect of each task will be provided in Section 5, on the basis of both statistical significance and machine learning experiments.

4. Machine Learning Experiments

So far we have described our extracted text-based features, and investigated the significance level of their values among various speaker group pairs for the different speaking tasks. In the next part of our study we will show that these attributes can also serve as a basis for an effective automatic discrimination among the three speaker groups (i.e., HC, subjects having MCI, and patients suffering from AD). That is, now we will perform machine learning experiments, using the extracted features.

4.1 Classification

We performed the classification experiments with the use of Support-Vector Machines (Schölkopf et al. 2001); we employed the libSVM implementation (Chang and Lin 2011). To avoid overfitting due to having a large number of meta-parameters, we utilized a linear kernel, with the complexity (C) value explored in the range $10^{\{-5,-4,\ldots,0,2\}}$. We treated each subject as one independent example. We then standardized each feature so as to have a zero mean and unit variance.

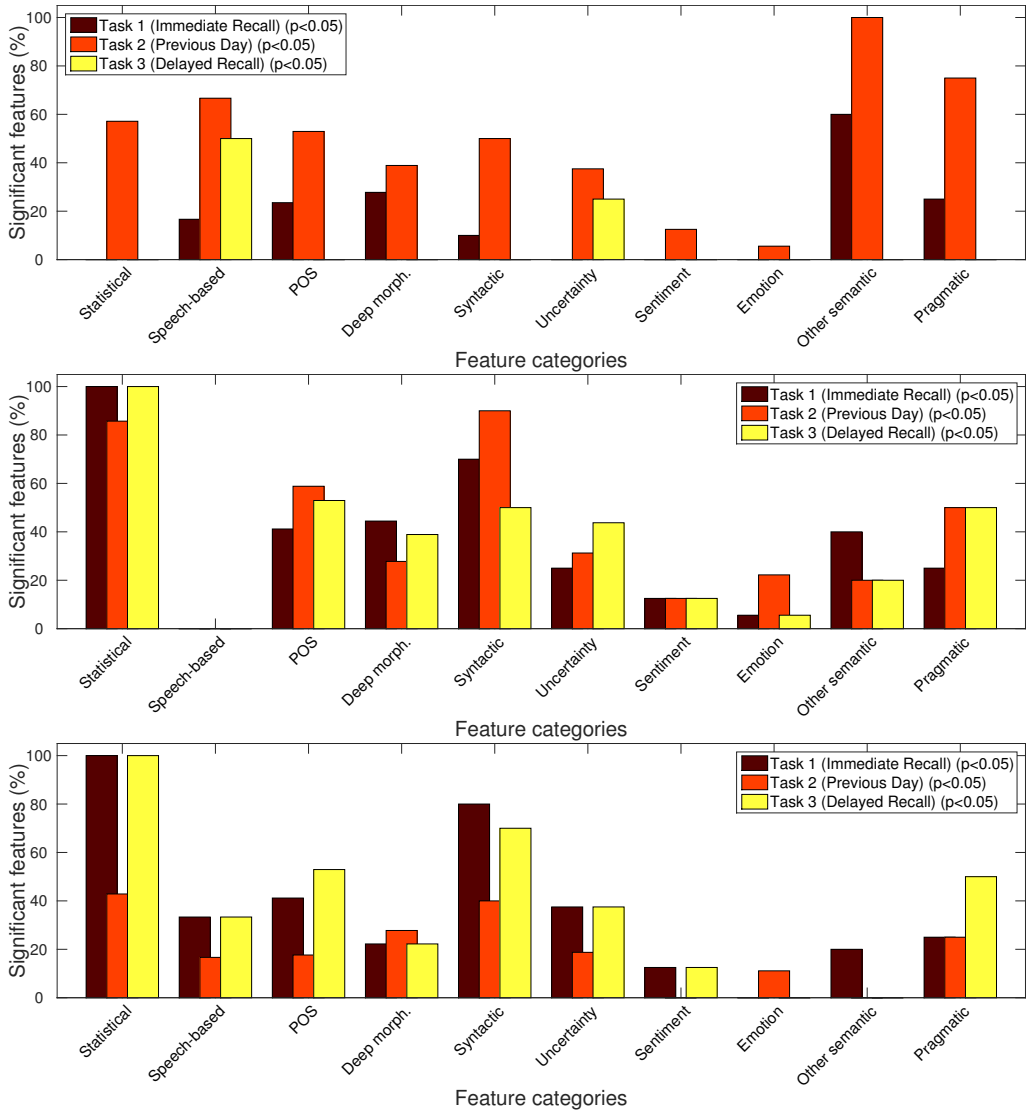

Figure 1
Ratio of attributes for the three speaker tasks and the feature categories examined, which significantly differ from $p < 0.05$ for the HC-MCI (top), MCI-mAD (middle), and HC-mAD (bottom) speaker categories.

From a machine learning perspective, having only 75 examples (i.e., subjects) is an extremely small dataset. However, the number of diagnosed MCI and mAD patients is limited; moreover, collecting and transcribing their speech and obtaining a medical diagnosis is time-consuming. Other similar studies we are aware of involved fewer than 100 patients (Satt et al. 2014; Jarrold et al. 2014; Lehr et al. 2012; Roark et al. 2011; Fraser, Rudzicz, and Rochon 2013; Weiner, Herff, and Schultz 2016). Having so few examples, we did not create separate training and test sets, but opted for cross-validation. In order to guarantee that each fold had the same number of speakers from each speaker group, we used 5-fold cross-validation: We divided the subjects into 5 groups (folds),

all containing 5 MCI and 5 mAD speakers, and 5 HC. Then we always trained on the features extracted from the speech of 60 speakers, from which 20 had MCI, 20 had mAD, and 20 were HC (i.e., 4 folds). Next, this machine learning model was evaluated on the remaining fifth fold (the data of 15 speakers), thereby guaranteeing that the same speaker's data was not used during training and evaluating the same machine learning model. Repeating this process for all folds, we obtained our predictions for all the 75 speakers. For comparison, we ran a baseline experiment, using only features that were proposed before (i.e., our novel features were excluded), but with the same settings mentioned above.

4.2 Evaluation

The choice of evaluation metric is not a clear-cut issue for this task. First of all, we can simply use the traditional classification accuracy score, since the class distribution is balanced for this dataset. However, besides indicating how well the subjects were identified as the members of each category, this task can also be viewed as a detection task, where we are interested in whether the speaker has *any* sort of cognitive disorder, that is, treating the MCI and mAD categories together as the positive class, while HC formed the negative class. As in this case the class distribution becomes imbalanced (25 control subjects and 50 subjects having some kind of cognitive disorder), we will also report (two-class) classification accuracy scores, but standard Information Retrieval metrics of *precision* and *recall* might also be useful. As there is evidently a trade-off between these two scores, they are usually aggregated together by the *F-measure* (or F_1-*score*), which is the harmonic mean of precision and recall. In the experiments we will present (3-class) accuracy scores and all the four 2-class scores (i.e., accuracy, precision, recall, and F-measure). As the last evaluation metric, we calculated the area under the ROC curve (AUC). We will report the AUC value of the HC class (reflecting how well the healthy subjects could be distinguished from either the MCI or the mAD speaker groups) as well as the unweighted mean of the AUC score for the three speaker categories. We tuned the meta-parameters (such as complexity of SVM) by choosing the one that led to the highest mean AUC value.

4.3 Handling the Three Tasks

Recall (see Section 2) that, in our recording setup, each subject performed three different tasks, leading to three different utterances. This means that the attributes we calculated (see Section 3.1) could be extracted from the transcripts of three different speech recordings, each one differing in the memory function triggered. In the simplest approach, the attributes calculated based on the three recordings were concatenated. Of course, because the three utterances differed by nature, we were also interested in the difference among these subject tasks. To this end, we also performed experiments using the features extracted from only one of these transcriptions.

4.4 Results

In our baseline experiment, we obtained an accuracy of 56% when identifying 3 classes of patients, with a precision of 0.556, a recall of 0.560, and an F-score of 0.557.

Table 4
Machine learning results obtained with the different linguistic attribute categories.

	3-class	2-class				AUC	
Features	**Accuracy**	**Accuracy**	**Precision**	**Recall**	F_1	**HC**	**mean**
Statistical	50.7%	62.7%	72.9%	70.0%	71.4	0.727	0.725
Speech	54.7%	64.0%	78.0%	64.0%	70.3	0.713	0.700
Morph. (all)	61.3%	76.0%	78.6%	88.0%	83.0	0.818	0.780
POS	57.3%	72.0%	78.4%	80.0%	79.2	0.743	0.734
Deep morph.	61.3%	74.7%	78.2%	86.0%	81.9	0.750	0.725
Syntactic	58.7%	72.0%	85.4%	70.0%	76.9	0.742	0.699
Semantic (all)	46.7%	65.3%	70.7%	82.0%	75.9	0.674	0.670
Uncertainty	48.0%	64.0%	71.7%	76.0%	73.8	0.690	0.671
Sentiment	42.7%	61.3%	71.4%	70.0%	70.7	0.574	0.527
Emotion	37.3%	52.0%	65.2%	60.0%	62.5	0.448	0.520
Other	34.7%	61.3%	69.8%	74.0%	71.8	0.569	0.558
Pragmatic	54.7%	72.0%	80.9%	76.0%	78.4	0.720	0.687
Demographic	41.3%	64.0%	78.0%	64.0%	70.3	0.708	0.585
All (w/o demogr.)	68.0%	77.3%	82.4%	84.0%	83.2	0.845	0.822
All (w. demogr.)	70.7%	80.0%	84.3%	86.0%	85.1	0.847	0.823

Table 4 shows the metric values we obtained for the various feature subsets. We can see that utilizing all the features led to actually quite competitive scores, either with or without the demographic information: the 68%–70% 3-class accuracy values, in our opinion, are quite high, and the two-class classification accuracy values of 77.3%–80% and the F_1-scores of 84–86 reflect a fine classification performance as well. These values also outperform our baseline results, hence the added value of our new features is justified. In the AUC values the difference was also even smaller: We measured values of 0.845–0.847 for the HC category, and the mean AUC of the three speaker groups was 0.822–0.823. This difference suggests that it was more straightforward to make a binary decision (i.e., whether the actual subject has *any* form of mental disorder) than to distinguish between the MCI and mAD categories, since we obtained lower AUC scores for the MCI and mAD classes than for the HC category.

Regarding the various attribute types, Table 4 displays the effectiveness of statistical features as an indicator of MCI and mAD: The relatively high scores (AUC values of 0.727 and 0.725, HC category and average, respectively) indicate that even with these simple descriptive features, dementia can be identified considerably above the level of chance. The semantic attributes, however, generally led to low scores. Uncertainty attributes seem to be the only exception (AUC values of 0.690 and 0.671), probably because of the difficulties in recalling things and events as the dementia becomes more and more progressive (see Section 3.2). Using just the pragmatic attributes, the classification results are moderate as well: The values (an accuracy of 72% and F_1-score of 78.4) are in clear contrast with the 3-class accuracy score of 54.7%, and although we achieved a fair AUC score for the HC speaker group (0.720), the mean AUC value of

0.687 suggests that the pragmatic attributes vary only slightly between the MCI and the mAD speaker groups.

4.5 Results Using the Significant Attributes Only

Next, we sought to fuse our previous experiments: We performed machine learning experiments, but this time we used only those attributes that showed a statistically significant difference. Filtering the attributes on the basis of statistical significance is a well-known feature selection method (see, e.g., Satt et al. 2013; Fraser, Rudzicz, and Rochon 2013; Kiss and Vicsi 2017; Tóth et al. 2018), which was reported to improve classification performance when detecting a wide variety of illnesses.

The experimental setup of our classification experiments matched that of our previous experiments. We kept only those attributes that had shown a statistically significant difference at the rate $p < 0.05$. In this way, we treated the three subject tasks as independent, that is, if an attribute was found to be significant only for the *immediate recall* subject task, we could have discarded the same attribute in the delayed recall and previous day tasks. However, the p-values were calculated for speaker group pairs; we selected the given attribute if it was found to be statistically significant for *any* of these pairs (e.g., HC and MCI).

Table 6 lists our results. In most cases, discarding the irrelevant (or at least, statistically not significant) attributes improved classification performance. This was especially the case when we utilized all attribute types (either with or without the demographic information): The mean AUC values improved from 0.822–0.823 to 0.847. Perhaps more important, the AUC value of the HC speaker group, which reflects how well we could tell whether the actual subject has *any* sort of mental illness, rose from 0.847 to 0.889, and from 0.845 to 0.891, when utilizing and when discarding the demographic attributes, respectively. As the three groups examined were matched for age, gender, and education, it is probable that this type of demographic information just confused the algorithm.

Regarding the statistical attributes, the evaluation metric values did not change at all, which is quite reasonable, since all such attributes were found to be significant with $p < 0.05$. Examining the performance of the classifier models trained on the speech-based features, we can see a large improvement in the 2-class case, as classification accuracy, precision, recall, and the F_1-score all rose by about 10% absolute (although the AUC values did not change significantly). Regarding morphological attributes, all values except recall improved notably: The F-measure value of 85.1 and the AUC value of 0.865 for the HC speaker category are, in our opinion, quite high. Examining the morphological attribute subtypes, this classification performance is mostly due to the deep morphological attributes, although using only the POS features led to high evaluation metric values as well. This is perhaps due to the morphologically rich nature of Hungarian.

Retaining only the significant syntactic attributes also led to nice improvements in four out of the seven scores; however, the AUC values of 0.754 and 0.739, HC speaker group and mean, respectively, are still among the lower ones obtained, indicating that these features are less useful for identifying dementia. On the other hand, relying on the semantic attributes was much more effective: When using all four feature subtypes, we achieved an F_1-score of 81.6 and an AUC score of 0.846 for the HC class. Clearly, this performance is due to the utility of the *uncertainty* feature subtype, as the remaining three groups (i.e., *sentiment*, *emotion*, and *other semantic*) in general led to rather

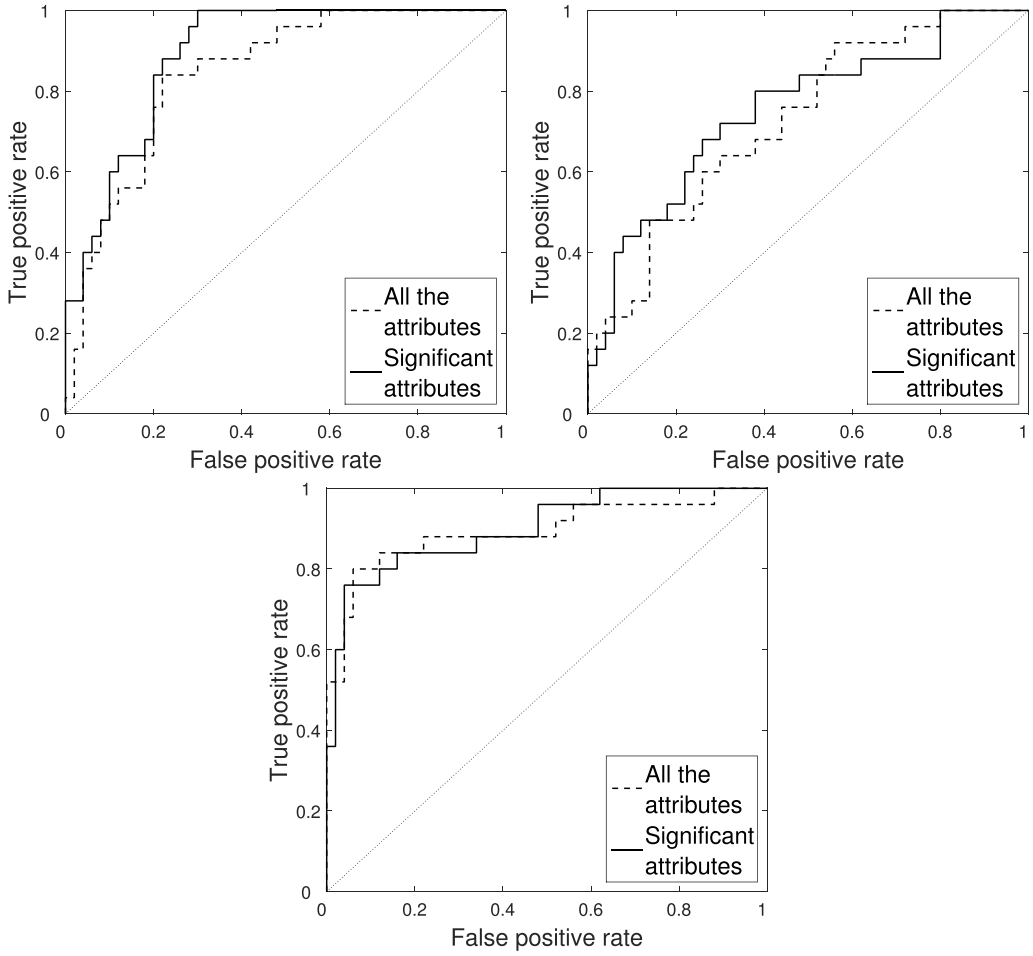

Figure 2
AUC curves for the HC (top left), MCI (top right), and mAD (bottom) speaker groups when using all attributes (excluding demographic ones) and when using only the attributes that were found to show a statistically significant difference with $p < 0.05$.

poor classification scores. Considering that MCI and mAD are both reported to cause difficulties in recalling things, this can be expected. Lastly, using the pragmatic attributes that were found to show statistically significant differences led the AUC value of the HC speaker category to increase from 0.720 to 0.743; still, this classification performance can be considered mediocre at best.

Figure 2 shows the AUC values for the three speaker categories when using all attributes (except demographic ones), and when using only the statistically significant ones. In the case of the HC speaker group (left side) the improvement from 0.845 to 0.891 brought by feature selection is clearly visible. Regarding the MCI group (middle), it is clear that this class was the hardest to identify, which is reasonable, as MCI is considered as the prodromal stage of AD, therefore the speech produced by these subjects differs only slightly from either the control subjects or those who already have dementia. Still, this graph demonstrates that using only the statistically significant attributes improved

Table 5
Results obtained for each of the tasks. IM: immediate recall, PD: previous day, DR: delayed recall, Acc.: accuracy, P: precision, R: recall.

		3-class	2-class				AUC	
Features	**Task**	**Acc.**	**Acc.**	**P**	**R**	F_1	**HC**	**mean**
Statistical	#IR	56.0%	70.7%	69.4%	100.0%	82.0	0.701	0.740
	#PD	46.7%	57.3%	76.5%	52.0%	61.9	0.642	0.688
	#DR	50.7%	65.3%	70.0%	84.0%	76.4	0.743	0.761
Speech-based	#IR	49.3%	62.7%	75.0%	66.0%	70.2	0.685	0.659
	#PD	42.7%	60.0%	77.8%	56.0%	65.1	0.667	0.610
	#DR	46.7%	65.3%	81.6%	62.0%	70.5	0.643	0.624
Morph. (all)	#IR	56.0%	68.0%	75.0%	78.0%	76.5	0.720	0.746
	#PD	46.7%	58.7%	74.4%	58.0%	65.2	0.651	0.629
	#DR	48.0%	60.0%	72.7%	64.0%	68.1	0.682	0.686
POS	#IR	54.7%	72.0%	76.4%	84.0%	80.0	0.730	0.715
	#PD	46.7%	62.7%	69.6%	78.0%	73.6	0.560	0.603
	#DR	53.3%	69.3%	72.9%	86.0%	78.9	0.685	0.702
Deep m.	#IR	60.0%	70.7%	74.1%	86.0%	79.6	0.641	0.700
	#PD	41.3%	53.3%	67.4%	58.0%	62.4	0.577	0.591
	#DR	46.7%	62.7%	72.9%	70.0%	71.4	0.654	0.646
Syntactic	#IR	57.3%	70.7%	70.6%	96.0%	81.4	0.734	0.744
	#PD	49.3%	60.0%	77.8%	56.0%	65.1	0.622	0.645
	#DR	49.3%	60.0%	72.7%	64.0%	68.1	0.703	0.722
Pragmatic	#IR	42.7%	56.0%	67.3%	66.0%	66.7	0.557	0.605
	#PD	45.3%	64.0%	73.5%	72.0%	72.7	0.634	0.585
	#DR	46.7%	61.3%	75.6%	62.0%	68.1	0.602	0.627
All (w/o dem.)	#IR	57.3%	68.0%	76.0%	76.0%	76.0	0.756	0.763
	#PD	50.7%	62.7%	77.5%	62.0%	68.9	0.732	0.669
	#DR	53.3%	66.7%	73.6%	78.0%	75.7	0.685	0.703
All (w. dem.)	#IR	58.7%	69.3%	76.5%	78.0%	77.2	0.760	0.782
	#PD	53.3%	65.3%	78.6%	66.0%	71.7	0.768	0.692
	#DR	52.0%	64.0%	72.5%	74.0%	73.3	0.674	0.699

the AUC value of this class from 0.726 to 0.750. Lastly, examining the AUC curves corresponding to the mAD subjects, we can note that the SVMs were able to identify these subjects with a high confidence (AUC score of 0.894); however, utilizing only the significant attributes could not improve the performance noticeably (AUC value of 0.901). In fact, this means that discarding the non-significant features helped the classifier model where it is the most useful: in distinguishing subjects having MCI from HC.

Table 6
Machine learning results obtained with the different linguistic attribute categories when using only attributes that displayed a statistically significant difference; cases where an improvement of at least 2% (accuracy, precision, recall, and F_1) or 0.02 (AUC) was observed are shown as **bold**.

	3-class	2-class				AUC	
Features	**Accuracy**	**Accuracy**	**Precision**	**Recall**	F_1	**HC**	**mean**
Statistical	50.7%	62.7%	72.9%	70.0%	71.4	0.727	0.725
Speech-based	**57.3%**	**74.7%**	**86.0%**	**74.0%**	**79.6**	0.698	0.706
Morph. (all)	**68.0%**	**80.0%**	**84.3%**	86.0%	**85.1**	**0.865**	**0.824**
POS	52.0%	73.3%	80.0%	80.0%	80.0	**0.825**	**0.760**
Deep morph.	**64.0%**	76.0%	79.6%	86.0%	82.7	**0.847**	**0.802**
Syntactic	**61.3%**	73.3%	82.6%	**76.0%**	**79.2**	0.754	**0.739**
Semantic (all)	**62.7%**	**76.0%**	**83.3%**	80.0%	**81.6**	**0.846**	**0.783**
Uncertainty	**54.7%**	**68.0%**	**78.3%**	72.0%	75.0	**0.748**	**0.724**
Sentiment	**50.7%**	**73.3%**	**75.9%**	**88.0%**	**81.5**	**0.623**	**0.606**
Emotion	36.0%	**58.7%**	66.1%	**78.0%**	**71.6**	**0.522**	**0.575**
Other	**37.3%**	56.0%	68.9%	62.0%	65.3	**0.623**	0.562
Pragmatic	56.0%	70.7%	**83.3%**	70.0%	76.1	**0.743**	0.701
Demographic	40.0%	61.3%	**81.8%**	54.0%	65.1	0.721	0.587
All (w/o demogr.)	69.3%	**80.0%**	84.3%	**86.0%**	85.1	**0.891**	**0.847**
All (w. demogr.)	65.3%	76.0%	79.6%	86.0%	82.7	**0.889**	**0.847**

5. Discussion

Now, we shall analyze the results in more detail and draw some conclusions about the relevance of each speaking task.

5.1 Analysis of the Effect of Feature Groups

As mentioned earlier, almost all features proved to be statistically significant when working with only two classes, that is, distinguishing only HC and patients with some kind of dementia. Hence, in the following we will focus on significant differences among the three groups (i.e., Tables 7–9), as we are interested in how the various groups of linguistic features may be affected as the disease progresses.

Upon analyzing the significance of **statistical features**, it was found that only Task 2 reveals differences among controls and MCI patients. However, all the tasks and almost all the features exhibit significant differences between the MCI and mAD group, which suggests that as these features deteriorate, patients tend to speak less and less as AD progresses. Hence, the diagnostic utility of statistical features can be fully exploited for differentiating the latter two groups.

As for the **speech features**, it is striking that there are no significant differences between the MCI and mAD groups (which is easily identifiable in Figure 1), but hesitations and pauses indicate significant changes among controls and MCI patients. Hence, speech features mostly define the border between these two groups, which suggests

that speech factors are already adversely affected in the early stage of dementia, making them good candidates for diagnostic purposes in order to detect dementia as early as possible. However, this group of features is less useful for distinguishing MCI and mAD patients.

Morphological features, especially the rates of nouns, verbs, pronouns, and adverbs, are good indicators of dementia in an early stage of the disease, in the case of Tasks 1 and 2. However, features in Task 3 (delayed recall) only exhibit significant differences in a later stage, that is, between MCI and mAD patients (see Figure 1). Thus, when the goal is to detect dementia as early as possible based on morphology, we should focus on Tasks 1 and 2.

Similar to the statistical features discussed above, the **syntactic abilities** of the speakers seem to decline over time as there is a higher number of significant differences among MCI and mAD patients, while only a few features distinguish controls and MCI patients (e.g., the number of subjects, objects, coordinations, and subordinations). Concerning the occurrence of coordinations and subordinations, we supposed that subordinate (dependent) clauses occur with higher frequency in the data of the control group. However, the rate of coordinations and subordinations led us to conclude that healthy controls do not tend to use more subordinate or coordinate clauses. Again, Task 3 seems to be relevant only in distinguishing the MCI and mAD classes.

Examining **semantic features**, we see that uncertainty features are responsible for most of the significant differences. This is especially true for epistemic and doxastic uncertainty (related to beliefs) and weasels (related to indefiniteness), which are of importance here. As dementia progresses, patients have difficulty in recalling things and events, hence the number of uncertain and fuzzy expressions like *someone*, *I think*, and so forth, increases. In spite of this, sentiment and emotion features in general did not prove to be effective in distinguishing the classes, only a few of these being significant for some groups, especially in Task 2. It should also be mentioned that whenever there is a significant difference, it is mostly related to negative sentiments and emotions such as anxiety and disgust. Even for positive emotions like joy and love, their number and rate decreases as dementia progresses. That is, it seems that patients with MCI and mAD express their thoughts in more negative ways than healthy controls do. Also, it should be noted that sentiment and emotion features in Tasks 1 and 3 tend to be significant mostly for the MCI–mAD distinction, which implies that these features are adversely affected in a later stage of the disease. However, other semantic features tend to be indicative of MCI, especially in Task 2, which means that when recalling the events of the previous day, MCI patients use significantly more phrases referring to memory activity, which is a clear indication of having memory problems. Also, the ratio of function words increases in their speech, that is, they may have difficulties with finding content words.

Lastly, among **pragmatic features** the discourse markers prove to be one of the most effective features. Discourse markers are special types of pragmatic markers that form part of an utterance, but they do not contribute to the meaning of the proposition per se (Fraser 2009). These lexical expressions are classified not syntactically, but in terms of their semantic/pragmatic functions. According to Fraser (2009), discourse markers basically signal a relation between the utterance which hosts them and the prior utterance. For instance: *you know*, *actually*, *basically*, *I mean*, or *so* in English or *mármint* 'I mean', *tudniillik* 'namely', *tudod* 'you know', *akkor* 'then', or *szerintem* 'in my opinion' in Hungarian. Based on the results of our analysis we may conclude that the more the disease progresses, the more likely the patient's speech will contain discourse markers.

Table 7
Significance of statistical and morphological features in the 3-class task. #: number, %: frequency, T1: immediate recall task, T2: previous day task, T3: delayed recall task.

	HC vs. MCI			MCI vs. mAD			HC vs. mAD		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
Statistical									
token#		**		**	**	***	**		***
sentence#				***	***	***	***	**	***
lemma#		**		**	**	***	**		***
lemma%		*		***		*	**		**
word#		**		***	**	***	***		***
sentence length				**	**	*	***	***	*
Morphological									
POS									
unknown#		*							
unknown%				*					
verb#		**		**	**	***	***		***
verb%	*						*		
noun#		*		**	**	**	***		***
noun%	**	**			*				
adjective#					*	***			***
adjective%							*		*
pronoun#		***		***	*	***	*		**
pronoun%	**					*		*	
conjunction#		**		*	**	**			*
conjunction%		*			*				
numeral#				**	***		**	*	
numeral%					*	*		**	*
adverb#		**			*	*			
adverb%	*						*		*
punctuation#		**			**	***			**
Deep morphological									
comparative#	*			*					
comparative%				*					
past tense#		**		**	**	*	*		**
past tense%						*			
present tense#		*			*	***			*
present tense%						*			
imperative verb#		**		*				*	
imperative verb%		**						*	
conditional verb#	*					*			*
conditional verb%	*						*		
Pl1 verb#		*		*			*	**	
Pl1 verb%					**			**	
Sg1 verb#	*	*		*	***	*		*	
demonstrative pronoun#		**		*	*	**			**
avg # of nominal suffixes	*			*			***		

In our machine learning experiments, we analyzed the efficacy of each feature group separately. We found that after analyzing all the tasks, statistical, morphological, and syntactic features proved to be the most useful (see Tables 7–12). Still, semantic features are less effective when used on their own, giving only an accuracy score of less

Table 8
Significance of syntactic and semantic features in the 3-class task. #: number, %: frequency, T1: immediate recall task, T2: previous day task, T3: delayed recall task.

	HC vs. MCI			MCI vs. mAD			HC vs. mAD		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
Syntactic features									
subject#		**		***	*	***	**		**
subject%				**	*		**	**	*
object#		**		**	**		***		
object%				**	*	*	**	**	*
subordination#		**		***	**	***	**		***
subordination%				*	**		**	**	
adverbial#		**			*	**			*
adverbial%	*						*		*
coordination#		**			*				
coordination%				**	**	*	**	**	*
Semantic features									
Uncertainty features									
uncertain#		**		*	**	**			**
uncertain%		*						*	
epistemic#					*	**	**		**
epistemic%					*	**	**	*	**
condition#		**	**		**	**			
condition%		*	*						
weasel#				**	*	*	*		*
weasel%		**		*		*	*	*	**
peacock#			*						
peacock%			*						
hedge#		*							
doxastic#				*		**	**		**
doxastic%							*		
Sentiment features									
negative#		*			*	*			*
positive%				*			*		
Emotion features									
love#				*					
anxiety#					*			*	
anxiety%					*				
disgust#		*							
joy%						*			
fear%								*	
emotive negative#					*				
emotive negative%					*				
Other semantic features									
memory%		*					*		
memory#		*				*			
negation word#	*	***			*				
content word%	*	**		**					
function word%	*	*		**					

than 50%. The same is true for the scenario of merging MCI and mAD patients (i.e., the 2-class identification task).

Morphological features seem to play an important role in machine learning experiments. Considering all the tasks, only by using morphological features, can we obtain an accuracy score of 61.33%, and when relying only on one of the tasks, high accuracy

Table 9
Significance of speech-based and pragmatic features in the 3-class task. #: number, %: frequency, T1: immediate recall task, T2: previous day task, T3: delayed recall task.

	HC vs. MCI			MCI vs. mAD			HC vs. mAD			
	T1	T2	T3	T1	T2	T3	T1	T2	T3	
Speech-based										
hesitation#		*	*							
hesitation%	*		*				**		*	
filled pause#		*	*							
pause#		*								
lengthened sound#		*						**		
pause after article#							*		*	
Pragmatic features										
speech act#		*				**			*	
discourse marker#		**		*	*	**				
discourse marker%	*	*			*		**	***	*	

scores can be again attained (i.e., 56%, 46.7%, and 48% for Task 1, 2, and 3, respectively—see Table 5). As the disease progresses, an impoverishment of morphology can be observed in the data: For instance, the number of verbs and nouns (and basically those of all parts of speech) decrease over time and the average number of nominal suffixes decreases with the progress of dementia. This might explain why morphological features are effective in separating the groups of speakers.

Uncertainty features exhibit significant differences among the groups, as well as being relevant in the machine learning experiments, especially in Task 3. As mentioned before, the reason for this might lie in the fact that dementia causes difficulties in recalling what happened earlier, meaning that speakers tend to express their uncertainty with linguistic cues too. Also, as Task 3 took place at the end of each recording session, speakers probably became tired by that time, resulting in a higher number of uncertainty cues.

5.2 Analysis of the Effect of the Tasks

Next, we would like to emphasize the strengths and weaknesses of each task, in order to determine which task is the most appropriate for identifying speakers with dementia.

When the tasks are considered separately (see Table 5), there are some interesting tendencies that should be examined further. For Task 1, statistical, morphological, and syntactic features are the most effective, but the role of emotion features is significant in the 2-class identification task, especially regarding recall. In Task 2, it is the sentiment features and other semantic features that have a positive effect on recall, and statistical and morphological features seem less important here. Moreover, uncertainty features prove to be effective in Task 2, together with morphological and statistical features. Overall, we may conclude that morphological and statistical features can perform well for all three tasks, while the efficacy of semantic features depends on the actual task.

The results for Tasks 1 and 2 indicate that semantic features can influence the results more strongly for the 2-class identification task than for the 3-class task. As expected,

Table 10
Significance of statistical and morphological features in the 2-class (HC vs. MCI/mAD) task. #: number. %: rate.

Statistical features	Imm.rec.	Prev.day	Del.rec.
token#	***	***	***
token%	***	***	*
sentence#	***	***	***
lemma#	***	***	***
lemma%	***	***	***
word#	***	***	***
sentence length	**	***	*
Morphological features			
POS features			
unknown%	***	***	***
verb%	*	**	**
noun#	***	***	***
noun%	***	***	***
adjective#	***	***	***
adjective%	***	***	***
pronoun#	***	**	***
pronoun%	***	***	***
conjunction#	*		***
conjunction%			*
numeral#	***	***	***
numeral%	***	***	***
adverb#	***	***	***
adverb%	***	***	***
punctuation#	***	***	***
Deep morphological features			
superlative#	**	**	*
superlative%	**	**	
comparative#			*
comparative%	***	***	***
sg1Pron#	***		***
past#	***		***
past%	***	***	***
present#	***	***	***
present%	*	***	
imperative#			**
imperative%			**
conditional verb#	**		***
conditional verb%		***	
pl1Verb#	***	***	***
pl1Verb%	***	***	***
demonstrative pronoun#	***	***	***
demonstrative pronoun%	***	***	**
average # of nominal suffixes	***	***	***

binary classification is an easier task to handle; it yields better scores for all feature groups, but it should also be added that a larger number of semantic features reveals significant differences in the 2-class task than in the 3-class task. This may mean that semantic features are more sensitive indicators of speakers with dementia, which is in accordance with our finding that semantics seems to be affected only in a later stage of the disease.

Table 11
Significance of semantic features in the 2-class (HC vs. MCI/mAD) task. #: number. %: rate.

Semantic features	Imm.rec.	Prev.day	Del.rec.
Uncertainty features			
uncertain#	***	***	***
uncertain%	***	**	***
epistemic#	***		***
epistemic%	***		***
investigation#	***	***	***
investigation%	***	***	***
condition#	***	***	***
condition%	***	***	***
weasel#	***	***	***
weasel%	***	***	***
peacock#	***	***	
peacock%	***	***	
hedge#		**	
doxastic#	***	*	***
doxastic%	**	***	**
hedge%		***	*
Emotion features			
joy#	*		*
joy%	*		*
fear#		**	
fear%		**	
anger#		**	
anger%		**	
love#			*
love%			*
surprise#	**	***	**
surprise%	***	***	***
sorrow%			*
Sentiment features			
positive#			*
positive%	*		*
negative%		**	**
slangPositive#		**	*
slangPositive%		*	***
slangNegative#	***	***	***
slangNegative%	***	**	**
negative emotive%	***	***	***
Other semantic features			
negation word#	**	*	**
content%	***	***	***
function%	***	***	**
memory#	***	*	***
memory%	***	***	***

It is also worth mentioning that Task 1 and Task 3 more effectively indicate the difference between the statistical features for the control group and the MCI and mAD patients. In connection with our previous experiences (see Sections 3.2 and 4.4), we may conclude that when the speakers have to tell a previously specified story (with given

Table 12
Significance of speech-based, syntactic, and pragmatic features in the 2-class (HC vs. MCI/mAD) task. #: number. %: rate.

Speech-based features	Imm.rec.	Prev.day	Del.rec.
hesitation#	***	***	***
hesitation%	***	***	***
filled pause#	***	***	***
pause#	**	***	**
lengthened sound#	*	***	***
pause after article#	***	***	**
Syntactic features			
subject#	***		***
subject%		**	
object#	***	***	***
object%	***	***	***
subordination#	***		***
subordination%	***	***	***
adverb#	***	***	***
adverb%	***	***	***
coordination#	***	***	***
coordination%	***	***	***
Pragmatic features			
speech act#	***	***	***
speech act%	***	***	***
discourse marker#	***	***	***
discourse marker%	***	***	***

content words, verbs, and story line) as in the case of Task 1 and Task 3, this restriction helps to highlight any mental disorder. However, in the case of Task 2 there is no such restriction so the topic, the content, and the order of the events are relatively free. The above-mentioned difference between the task types could possibly lead to the diverging frequency of parts-of-speech of the words as well.

In the machine learning experiments, it can be seen that in Tasks 1 and 3, the application of only morphological features results in a higher accuracy than applying all features. This is probably due to the fact that most semantic features perform poorly in these tasks—with the exception of uncertainty features in Task 3—which might harm performance. It is also interesting that in Task 1, statistical features can yield about the same accuracy (and even higher F-score) than morphological features in the 2-class identification task. Thus, it may be concluded that when our goal is to distinguish healthy controls from patients with dementia, it might be sufficient to rely on very simple statistical features in the immediate recall task.

In Task 2 (previous day), it is notable that the other semantic features behave very differently in the 2-class and the 3-class identification tasks. Namely, the use of only the other semantic features yields the best accuracy (66.67%) and the best F-score (80%) for the 2-class task but they are not useful for telling apart the 3 classes (cf. the accuracy score of 33.33%). Sentiment features exhibit a similar trend here: They achieve high accuracy scores in the 2-class task but only a lower accuracy score in the 3-class task. Hence, it is recommended that these types of features can effectively identify

people with dementia, but they are not sensitive enough to detect the subtle differences between the MCI and mAD groups.

Based on the statistical significance tests and our machine learning experiments, the following can be concluded with regard to each task type.

For Task 1 (immediate recall), statistical features exhibit significant differences among the MCI and mAD groups. The same is true for syntactic features. Also, when focusing on semantics, whenever we can find a statistically significant feature, it is related to the distinction of the MCI-mAD classes. In spite of this, morphological features can exhibit statistically significant differences for controls and speakers with dementia. In the machine learning experiments, deep morphological features seem to be the most effective in both the 2-class and the 3-class identification tasks; however, statistical and syntactic features also result in high accuracy and F-scores. In summary, this means that the strongest point of the immediate recall task is to distinguish the MCI and mAD groups. Moreover, when the goal is to identify speakers with dementia (i.e., no distinction among MCI and mAD speakers), it is sufficient to use only statistical features, without the need for any deep linguistic analysis, which makes it a very cost-effective procedure in the case where there is a short video at our disposal to play for the patients in the data collecting sessions.

In Task 2 (previous day), however, the other semantic features tend to achieve the highest F-score, and they also perform well in the 2-class identification task. As regards the significance of the features, statistical features are also strong here, as well as morphological, syntactic, and the other semantic features for both the distinction among control vs. MCI speakers and MCI vs. mAD speakers. Hence, the other semantic features tend to be distinctive for Task 2: controls use significantly more content words and fewer function words (such as conjunctions, articles, etc.) than speakers with dementia and they also use fewer phrases related to memory activity. To sum up, Task 2 does not require any specific preparation because it is based on a single question (*Tell me what happened yesterday*); however, deeper linguistic analysis is needed to profit from the distinctive features of this task.[1]

Lastly, Task 3 (delayed recall) seems to indicate the fewest number of significant differences among the control and MCI groups. This might be related to the fact that by the end of the session, speakers were tired and could not concentrate as well, hence it was difficult to find any differences in their cognitive abilities. Nevertheless, there are significant differences, for instance, for the statistical, morphological, and syntactic features, between the MCI and mAD groups. Moreover, if we consider our experiments with the three tasks, it is Task 3 where the overall highest F-score is the lowest, that is, the other two tasks can perform better in the machine learning experiments, although the difference is not considerable. The added value of Task 3 lies in distinguishing MCI and mAD, which justifies its inclusion in the experimental setup. To summarize, we can conclude that whenever we need a fine-grained distinction (i.e., distinguishing healthy controls, MCI, and mAD speakers), then the use of the immediate recall and the delayed recall tasks are strongly recommended (in addition to the previous day task).

Regarding the usefulness of each task, we performed one last machine learning experiment. We trained 3-class SVMs using only the attributes found to be significant (with $p < 0.05$), but using only the attributes corresponding to one of the speaker tasks

1 It needs to be added that the day of the visit might slightly influence the semantic content of the patient's utterances in this task. However, our feature set does not primarily focus on the semantic content; rather, the emphasis is on deeper linguistic features, which are probably independent of the semantic content or real-life activities in most cases.

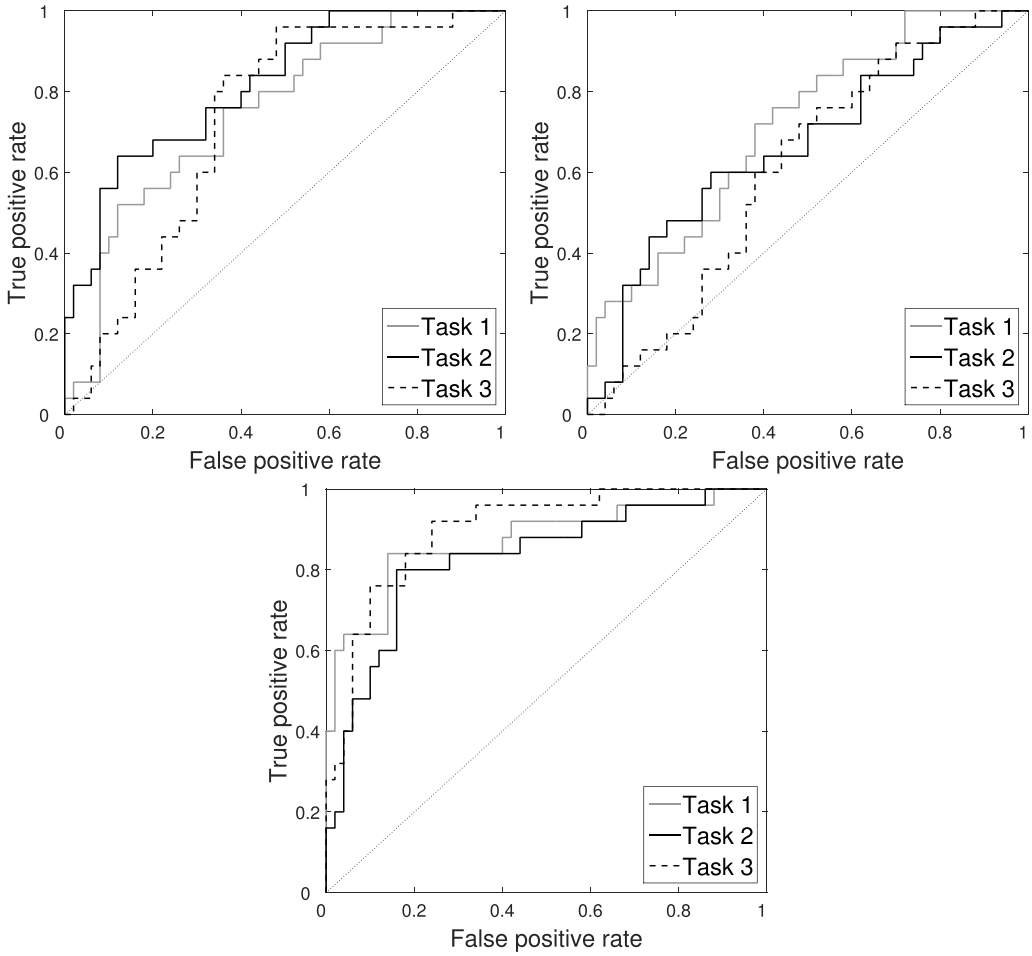

Figure 3
AUC curves for the HC (top left), MCI (top right), and mAD (bottom) speaker groups when using only the attributes extracted from one speaker task.

(again excluding demographic information). Figure 3 shows the measured AUC values for the three speaker categories. Of course, the AUC scores appeared to be lower than in the previous case, but our aim here was to focus on the usefulness of the different speaker tasks. Examining the AUC scores corresponding to the control subjects (see the left side of Figure 3), it is clear that the second task (i.e., *previous day*) contributed the most to the identification of these speakers (AUC score of 0.818), while the two recall tasks were noticeably less useful (AUC values of 0.748 and 0.726, Task 1 and Task 3, respectively). For the MCI speakers (see the middle of Figure 3), Task 1 (i.e., *immediate recall*) was found to be the most useful with an AUC score of 0.713, followed by Task 2 (*previous day*, AUC of 0.664), and, surprisingly, Task 3 (*delayed recall*) proved to be the worst one (AUC of 0.607). Regarding the subjects suffering from mAD (see the right side of Figure 3), all three tasks led to a high-quality identification of these subjects (AUC scores of 0.872, 0.828, and 0.898). Our hypothesis is that Task 2 is less useful in differentiating between MCI and mAD, which also contributed to its mediocre AUC

value for the MCI group; however, perhaps the most important aspect is to separate subjects having MCI from the healthy speakers, for which Task 2 (i.e., asking the subjects about their previous day) is the most useful.

6. Conclusions

In this article, we presented our methods for automatically identifying Hungarian patients suffering from MCI or mAD based on their speech transcripts. In our study, we utilized the Hungarian MCI–mAD database, recorded at the Memory Clinic at the Department of Psychiatry or the University of Szeged, Hungary. Here, we used 225 recordings performed by the subjects in three different tasks (immediate recall, delayed recall, and telling some words about the previous day).

In our experiments, we used a rich feature set (altogether 330 features) derived from the transcripts and the results of the automatic linguistic analyses performed with `magyarlanc`. We described each feature category in detail, then we presented the results of the statistical analysis of the data. We concluded that there are notable differences in the usability of not just the features, but also the speaker tasks as an indicator to differentiate between each group (i.e., HC, those with MCI, and those with mAD), as well.

In the next part of the study we showed how the various attributes can serve as a basis for an effective automatic discrimination among the three speakers groups. Our system used machine learning techniques on the basis of a rich feature set including parameters of linguistic characteristics of spontaneous speech as well as features exploiting morphological and syntactic parsing and semantic and pragmatic features. We concluded that, utilizing all features led to competitive scores, either with or without the demographic information (3-class accuracy scores: 68%–70%, 2-class classification accuracy scores: 77.3%–80%, F_1-scores: 84–86). In the AUC values the difference was even smaller (for the healthy control category: 0.845–0.847, for the three speaker groups: 0.822–0.823). This difference suggests that it is more straightforward to make a binary decision (i.e., whether the actual individual has *any* form of mental disorder) than to distinguish between the MCI and mAD categories.

Regarding the various attribute types, the analysis of the statistical differences indicate that even with these simple descriptive features, dementia can be identified notably above chance level. The semantic attributes, however, generally led to low scores, with uncertainty attributes being the only exception. Using only the pragmatic attributes, the results suggest that the pragmatic attributes vary just slightly between the MCI and the mAD speaker groups.

We also examined how the different data recording scenarios affect linguistic features, and concluded that when the goal is to distinguish MCI and mAD patients from healthy controls, the use of immediate recall and delayed recall tasks is strongly advisable, in addition to the previous day task.

In the future, we would like to extend our data set with new transcripts. Also, on the basis of the promising research results concerning some of the deep morphological, semantic, and pragmatic features, we will investigate whether combining certain sets of features can further improve the automatic detection of MCI and mAD.

Acknowledgments

This study was partially funded by the National Research, Development, and Innovation Office of Hungary via contract NKFIH FK-124413, by grant NKFIH-1279-2/2020 of the Hungarian Ministry of Innovation and Technology, and by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence

National Laboratory Program (MILAB). This work was supported by the Hungarian Research Fund (NKFIH / OTKA, grant number PD 132312). Gábor Gosztolya was also funded by the János Bolyai Scholarship of the Hungarian Academy of Sciences and by the Hungarian Ministry of Innovation and Technology New National Excellence Program ÚNKP-21-5-SZTE.

References

Al-Hameed, Sabah, Mohammed Benaissa, and Heidi Christensen. 2017. Detecting and predicting Alzheimer's Disease severity in longitudinal acoustic data. In *Proceedings of the International Conference on Bioinformatics Research and Applications 2017*, ICBRA 2017, pages 57–61. https://doi.org/10.1145/3175587.3175589

Al-Hameed, Sabah, Mohammed Benaissa, Heidi Christensen, Bahman Mirheidari, Daniel Blackburn, and Markus Reuber. 2019. A new diagnostic approach for the identification of patients with neurodegenerative cognitive complaints. *PLoS ONE*, 14(5):1–18. https://doi.org/10.1371/journal.pone.0217388, PubMed: 31125389

APA. 2000. *DSM-IV-TR*, American Psychiatric Association.

Balakrishnan, Vimala and Ethel Lloyd-Yemoh. 2014. Stemming and lemmatization: A comparison of retrieval performances. In *Proceedings of SCEI Seoul Conferences*, pages 174–179. https://doi.org/10.7763/LNSE.2014.V2.134

Baldas, Vassilis, Charalampos Lampiris, Christos N. Capsalis, and Dimitrios Koutsouris. 2010. Early diagnosis of Alzheimer's type dementia using continuous speech recognition. In *Proceedings of MobiHealth*, pages 105–110. https://doi.org/10.1007/978-3-642-20865-2_14

Boise, Linda, Margaret B. Neal, and Jeffrey Kaye. 2004. Dementia assessment in primary care: Results from a study in three managed care systems. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 59(6):M621–M626. https://doi.org/10.1093/gerona/59.6.M621, PubMed: 15215282

Bucks, R. S., S. Singh, J. M. Cuerden, and G. K. Wilcock. 2000. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91. https://doi.org/10.1080/026870300401603

Chang, Chih Chung and Chih-Jeh Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:1–27. https://doi.org/10.1145/1961189.1961199

Chapman, Sandra Bond, Jennifer Zientz, Myron Weiner, Roger Rosenberg, William Frawley, and Mary Hope Burns. 2002. Discourse changes in early Alzheimer disease, mild cognitive impairment, and normal aging. *Alzheimer Disease & Associated Disorders*, 16(3):177–186. https://doi.org/10.1097/00002093-200207000-00008, PubMed: 12218649

Croot, Karen, John R. Hodges, John Xuereb, and Karalyn Patterson. 2000. Phonological and articulatory impairment in Alzheimer's disease: A case series. *Brain and Language*, 75(2):277–309. https://doi.org/10.1006/brln.2000.2357, PubMed: 11049669

Dér, Csilla Ilona and Alexandra Markó. 2007. A magyar diskurzusjelölők szupraszegmentális jelöltsége, *Nyelvelmélet–nyelvhasználat*. Tinta, Székesfehérvár–Budapest, pages 61–67.

dos Santos, Leandro B., Edilson Anselmo Corrêa Jr., Osvaldo N. Oliveira Jr., Diego R. Amancio, Letícia L. Mansur, and Sandra M. Aluísio. 2017. Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. In *Proceedings of ACL*, pages 1284–1296. https://doi.org/10.18653/v1/P17-1118

Folstein, M. F., S. E. Folstein, and P. R. McHugh. 1975. Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198.

Fraser, Bruce. 2009. An account of discourse markers. *International Review of Pragmatics*, 1(2):293–320. https://doi.org/10.1163/187730909X12538045489818

Fraser, Kathleen C., Kristina Lundholm Fors, and Dimitrios Kokkinakis. 2018. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer, Speech & Language*, 53:121–139. https://doi.org/10.1016/j.csl.2018.07.005

Fraser, Kathleen C., Kristina Lundholm Fors, Dimitrios Kokkinakis, and Arto Nordlund. 2017. An analysis of eye-movements

during reading for the detection of mild cognitive impairment. In *Proceedings of EMNLP*, pages 1027–1037. https://doi.org/10.18653/v1/D17-1107

Fraser, Kathleen C., Frank Rudzicz, and Elizabeth Rochon. 2013. Using text and acoustic features to diagnose progressive aphasia and its subtypes. In *INTERSPEECH*, pages 2177–2181. https://doi.org/10.21437/Interspeech.2013-514

Freedman, M., L. Leach, E. Kaplan, G. Winocur, K. I. Shulman, and D. Delis. 1994. *Clock Drawing: A Neuropsychological Analysis*. New York: Oxford University Press.

Galvin, James E. and Carl H. Sadowsky. 2012. Practical guidelines for the recognition and diagnosis of dementia. *The Journal of the American Board of Family Medicine*, 25(3):367–382. https://doi.org/10.3122/jabfm.2012.03.100181, PubMed: 22570400

Garrard, P., L. M. Maloney, J. R. Hodges, and K. Patterson. 2005. The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, 128(2):250–260. https://doi.org/10.1093/brain/awh341, PubMed: 15574466

Gosztolya, Gábor, László Tóth, Tamás Grósz, Veronika Vincze, Ildikó Hoffmann, Gréta Szatlóczki, Magdolna Pákáski, and János Kálmán. 2016. Detecting Mild Cognitive Impairment from spontaneous speech by correlation-based phonetic feature selection. In *Proceedings of Interspeech*, pages 107–111. https://doi.org/10.21437/Interspeech.2016-384

Gosztolya, Gábor, Veronika Vincze, László Tóth, Magdolna Pákáski, János Kálmán, and Ildikó Hoffmann. 2019. Identifying Mild Cognitive Impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. *Computer, Speech & Language*, 53(Jan):181–197. https://doi.org/10.1016/j.csl.2018.07.007

Hirst, Graeme and Vanessa Wei Feng. 2012. Changes in style in authors with Alzheimer's disease. *English Studies*, 93(3):357–370. https://doi.org/10.1080/0013838X.2012.668789

Hoffmann, Ildikó, Dezső Németh, Cristina D. Dye, Magdolna Pákáski, Tamás Irinyi, and János Kálmán. 2010. Temporal parameters of spontaneous speech in Alzheimer's disease. *International Journal of Speech-Language Pathology*, 12(1):29–34. https://doi.org/10.3109/17549500903137256, PubMed: 20380247

Holmes, David I. and Sameer Singh. 1996. A stylometric analysis of conversational speech of aphasic patients. *Literary and Linguistic Computing*, 11(3):133–140. https://doi.org/10.1093/llc/11.3.133

Janka, Z., A. Somogyi, E. Maglóczky, Magoldna Pákáski, and János Kálmán. 1988. Dementia szűrővizsgálat cognitív gyorsteszt segítségével. *Orvosi hetilap*, 129:297–299.

Jarrold, William, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria L. Gorno-Tempini, and Jennifer Ogar. 2014. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 27–37.

Kálmán, János, E. Maglóczky, and Z. Janka. 1995. Óra Rajzolási Teszt: Gyors és egyszerű dementia szűrőmódszer. *Psychiatria Hungarica*, 10(3):11–18.

Kiss, Gábor and Klára Vicsi. 2017. Mono- and multi-lingual depression prediction based on speech processing. *International Journal of Speech Technology*, 20(4):919–935. https://doi.org/10.1007/s10772-017-9455-8

Kokkinakis, Dimitrios, Kristina Lundholm Fors, Eva Björkner, and Arto Nordlund. 2017. Data collection from persons with mild forms of cognitive impairment and healthy controls – infrastructure for classification and prediction of dementia. In *Proceedings of NoDaLiDa*, pages 172–182.

König, Alexandra, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, P. Robert, and Renaud David. 2015. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):112–124. https://doi.org/10.1016/j.dadm.2014.11.012, PubMed: 27239498

Kutuzov, Andrey and Elizaveta Kuzmenko. 2019. To lemmatize or not to lemmatize: How word normalisation affects ELMo performance in word sense disambiguation. *arXiv preprint arXiv:1909.03135*.

Le, Xuan, Ian Lancashire, Graeme Hirst, and Regina Jokel. 2011. Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three British novelists. *Literary and Linguistic Computing*, 26(4):435–461. https://doi.org/10.1093/llc/fqr013

Lehr, Maider, Emily Prud'hommeaux, Izhak Shafran, and Brian Roark. 2012. Fully automated neuropsychological assessment for detecting Mild Cognitive Impairment. In *Proceedings of Interspeech*, pages 1039–1042. https://doi.org/10.21437/Interspeech.2012-306

Liu, Bing. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

López-de-Ipiña, K., J. B. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, C. M. Travieso, M. Ecay-Torres, P. Martínez-Lage, and H. Eguiraun. 2015. On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature. *Cognitive Computation*, 7(1):44–55. https://doi.org/10.1007/s12559-013-9229-9

López-de-Ipiña, Karmele, Jesus-Bernardino Alonso, Carlos Manuel Travieso, Jordi Solé-Casals, Harkaitz Egiraun, Marcos Faundez-Zanuy, Aitzol Ezeiza, Nora Barroso, Miriam Ecay-Torres, Pablo Martinez-Lage, and Unai Martinez de Lizardui. 2013. On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. *Sensors*, 13(5):6730–6745. https://doi.org/10.3390/s130506730 PubMed: 23698268

López-de-Ipiña, Karmele, Jordi Solé i Casals, Harkaitz Eguiraun, Jesús B. Alonso, Carlos Manuel Travieso-González, Aitzol Ezeiza, Nora Barroso, Miriam Ecay, Pablo Martinez-Lage, and Blanca Beitia. 2015. Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach. *Computer, Speech & Language*, 30(1):43–60. https://doi.org/10.1016/j.csl.2014.08.002

Lunsford, Rebecca and Peter A. Heeman. 2015. Using linguistic indicators of difficulty to identify mild cognitive impairment. In *Proceedings of Interspeech*, pages 658–662.

Matsuda, Hiroshi, Takashi Asada, and Aya Midori Tokumaru. 2017. *Neuroimaging Diagnosis for Alzheimer's Disease and Other Dementias*. Springer. https://doi.org/10.21437/Interspeech.2015-235

Mirheidari, Bahman, Daniel Blackburn, Kirsty Harkness, Traci Walker, Annalena Venneri, Markus Reuber, and Heidi Christensen. 2017. Toward the automation of diagnostic conversation analysis in patients with memory complaints. *Journal of Alzheimer's Disease*, 58(2):373–387. https://doi.org/10.3233/JAD-160507, PubMed: 28436388

Mirheidari, Bahman, Daniel Blackburn, Markus Reuber, Traci Walker, and Heidi Christensen. 2016. Diagnosing people with dementia using automatic conversation analysis. In *Proceedings of Interspeech*, pages 1220–1224. https://doi.org/10.21437/Interspeech.2016-857

Mladenović, Miljana, Jelena Mitrović, Cvetana Krstev, and Duško Vitas. 2016. Hybrid sentiment analysis framework for a morphologically rich language. *Journal of Intelligent Information Systems*, 46(3):599–620. https://doi.org/10.1007/s10844-015-0372-5

Nakata, Yasuhiro, Noriko Sato, Kiyotaka Nemoto, Osamu Abe, Shoko Shikakura, Kunimasa Arima, Nobuo Furuta, Masatake Uno, Shigeo Hirai, Yoshitaka Masutani, Kuni Ohtomo, A. James Barkovich, and Shigeki Aoki. 2009. Diffusion abnormality in the posterior cingulum and hippocampal volume: Correlation with disease progression in Alzheimer's disease. *Magnetic Resonance Imaging*, 27(3):347–354. https://doi.org/10.1016/j.mri.2008.07.013, PubMed: 18771871

Nelson, Lucy and Naji Tabet. 2015. Slowing the progression of Alzheimer's disease; what works? *Ageing Research Reviews*, 23(B):193–209. https://doi.org/10.1016/j.arr.2015.07.002, PubMed: 26219494

Patocskai, A. T., Magdolna Pákáski, G. Vincze, M. Fullajtár, Irma Szimjanovszki, K. Boda, Z. Janka, and János Kálmán. 2014. Is there any difference between the findings of clock drawing tests if the clocks show different times? *International Journal of Geriatric Psychiatry*, 39(4):749–757. https://doi.org/10.3233/JAD-131313, PubMed: 24270210

Petersen, R. C. C. 2004. Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, 256(3):183–194. https://doi.org/10.1111/j.1365-2796.2004.01388.x, PubMed: 15324362

Pistono, Aurélie, Jeremie Pariente, C. Bézy, B. Lemesle, J. Le Men, and Mélanie Jucla. 2019. What happens when nothing happens? An investigation of pauses as a compensatory mechanism in early Alzheimer's disease. *Neuropsychologia*, 124:133–143. https://doi.org/10.1016/j.neuropsychologia.2018.12.018, PubMed: 30593773

Roark, B., M. Mitchell, J. Hosom, K. Hollingshead, and J. Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *Transactions on Audio, Speech, and Language Processing*, 19(7):2081–2090. https://doi.org/10.1109/TASL.2011.2112351, PubMed: 22199464

Rosen, W. G., R. C. Mohs, and K. L. Davis. 1984. A new rating scale for Alzheimer's disease. *Journal of Psychiatric Research*, 141(11):1356–1364. https://doi.org/10.1176/ajp.141.11.1356, PubMed: 6496779

Satt, Aharon, Ron Hoory, Alexandra König, Pauline Aalten, and Philippe H. Robert. 2014. Speech-based automatic and robust detection of very early dementia. In *15th Annual Conference of the International Speech Communication Association*, pages 2538–2542. https://doi.org/10.21437/Interspeech.2014-544

Satt, Aharon, Alexandra Sorin, Orith Toledo-Ronen, Oren Barkan, Ioannis Kompatsiaris, Athina Kokonozi, and Magda Tsolaki. 2013. Evaluation of speech-based protocol for detection of early-stage dementia. In *Proceedings of Interspeech*, pages 1692–1696. https://doi.org/10.21437/Interspeech.2013-32

Scheltens, Philip, Nick Fox, Frederik Barkhof, and Charles De Carli. 2002. Structural magnetic resonance imaging in the practical assessment of dementia: Beyond exclusion. *Lancet Neurology*, 1(1):13–21. https://doi.org/10.1016/S1474-4422(02)00002-9

Schölkopf, Bernhard, John C. Platt, John Shawe-Taylor, Alexander J. Smola, and Robert C. Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471. https://doi.org/10.1162/089976601750264965, PubMed: 11440593

Shibata, Daisaku, Shoko Wakamiya, and Eiji Aramak. 2016. Detecting Japanese patients with Alzheimer's disease based on word category frequencies. In *Proceedings of ClinicalNLP*, pages 78–85.

Stricker, N. H., B. C. Schweinsburg, L. Delano-Wood, C. E. Wierenga, K. J. Bangen, K. Y. Haaland, L. R. Frank, D. P. Salmon, and M. W. Bondi. 2009. Decreased white matter integrity in late-myelinating fiber pathways in Alzheimer's disease supports retrogenesis. *Neuroimage*, 45(1):10–16. https://doi.org/10.1016/j.neuroimage.2008.11.027, PubMed: 19100839

Szabó, Martina Katalin. 2015, Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai és dilemmái. In *Segédkönyvek a nyelvészet tanulmányozásához 177*. Tinta, Budapest, pages 278–285.

Szabó, Martina Katalin, Orsolya Ring, Balázs Nagy, László Kiss, Júlia Koltai, Gábor Berend, László Vidács, Attila Gulyás, and Zoltán Kmetty. 2020. Exploring the dynamic changes of key concepts of the Hungarian socialist era with natural language processing methods. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 54(1):1–13. https://doi.org/10.1080/01615440.2020.1823289

Szabó, Martina Katalin, Veronika Vincze, and Gergely Morvay. 2016, Magyar nyelvű szövegek emócióelemzésének elméleti nyelvészeti és nyelvtechnológiai problémái. In *Távlatok a mai magyar alkalmazott nyelvészetben*. Tinta, Budapest, pages 282–292.

Taler, Vanessa and N. A. Phillips. 2008. Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. *Journal of Clinical and Experimental Neuropsychology*, 30(5):501–556. https://doi.org/10.1080/13803390701550128, PubMed: 18569251

Thomas, Calvin, Vlado Kešelj, Nick Cercone, Kenneth Rockwood, and Elissa Asp. 2005. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In *Mechatronics and Automation, 2005 IEEE International Conference*, volume 3, pages 1569–1574.

Tóth, László, Gábor Gosztolya, Veronika Vincze, Ildikó Hoffmann, Gréta Szatlóczki, Edit Biró, Fruzsina Zsura, Magdolna Pákáski, and János Kálmán. 2015. Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In *16th Annual Conference of the International Speech Communication Association*, pages 2694–2698. https://doi.org/10.21437/Interspeech.2015-568

Tóth, László, Ildikó Hoffmann, Gábor Gosztolya, Veronika Vincze, Gréta Szatlóczki, Zoltán Bánréti, Magdolna Pákáski, and János Kálmán. 2018. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Current Alzheimer Research*, 15(2):130–138. https://doi.org/10.2174/1567205014666171121114930, PubMed: 29165085

Vincze, Veronika. 2014. Uncertainty detection in Hungarian texts. In *Proceedings of Coling*, pages 1844–1853.

Vincze, Veronika, Gábor Gosztolya, László Tóth, Ildikó Hoffmann, Gréta Szatlóczki, Zoltán Bánréti, Magdolna Pákáski, and János Kálmán. 2016. Detecting Mild Cognitive Impairment by exploiting linguistic information from transcripts. In *Proceedings of ACL*, pages 181–187. https://doi.org/10.18653/v1/P16-2030

Weiner, Jochen, Christian Herff, and Tanja Schultz. 2016. Speech-based detection of Alzheimer's disease in conversational German. In *Proceedings of Interspeech*, pages 1938–1942. https://doi.org/10.21437/Interspeech.2016-100

Yin, Changhao, Siou Li, Weina Zhao, and Jiachun Feng. 2013. Brain imaging of mild cognitive impairment and Alzheimer's disease. *Neural Regeneration Research*, 8(5):435–444.

Zimny, A., P. Szewczyk, E. Trypka, R. Wojtynska, L. Noga, J. Leszek, and M. Sasiadek. 2011. Multimodal imaging in diagnosis of Alzheimer's disease and amnestic mild cognitive impairment: Value of magnetic resonance spectroscopy, perfusion, and diffusion tensor imaging of the posterior cingulate region. *Journal of Alzheimer's Disease*, 27(3):435–444. https://doi.org/10.3233/JAD-2011-110254, PubMed: 21841260

Zsibrita, János, Veronika Vincze, and Richárd Farkas. 2013. magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP*, pages 763–771.