Lexical Density in Slovak Speech: A Non-invasive Indicator for Alzheimer's Disease and Mild Cognitive Impairment

Nataliia Časnochova Zozuk, Lívia Kelebercová, and Daša Munková

Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, SK, 949 01 Nitra, Slovakia
nataliia.casnochova.zozuk@ukf.sk
lkelebercova@ukf.sk
dmunkova@ukf.sk

Abstract. Background: Alzheimer's disease (AD) and mild cognitive impairment (MCI) are increasingly prevalent neurodegenerative conditions that impact cognitive and linguistic functions. Early detection is crucial for timely intervention, yet traditional diagnostic methods, such as neuroimaging and cerebrospinal fluid analysis, are invasive and costly. This study investigates whether linguistic markers, specifically lexical density in spoken language, can serve as reliable, noninvasive indicators of cognitive impairment. **Methods:** Using data from 214 participants (healthy (CN), MCI, and AD) collected via a picture description task, we applied natural language processing (NLP) techniques to analyze various measures lexical density (*V1*/*W*, *R2*, *ADJ*/*W*, *ADV*/*W*). **Results:** Statistical analysis revealed significant associations between certain lexical metrics and cognitive impairment levels. Specifically, measures such as *V1*/*W* and *H1* differentiated healthy participants from those with MCI, while ADJ/W and ADV/W were particularly effective in distinguishing AD from cognitively normal participants. **Conclusion:** These findings suggest that linguistic features, related to lexical density, can provide insights into cognitive health and may offer a valuable tool for early detection of AD and MCI. Future research should explore broader linguistic metrics and additional language tasks to enhance diagnostic accuracy and facilitate the development of automated, speech-based screening tools.

Keywords: Lexical density, Mild cognitive impairment, Alzheimer's disease.

1 Introduction

Alzheimer's disease (AD) is the most common form of dementia and is expected to affect an increasing number of people each year as life expectancy rises [1-2]. Characterized by irreversible neuronal loss, particularly in the cortex and hippocampus, AD results in progressive memory impairment (or loss) and behavioral changes and generally follows a preliminary stage known as mild cog-

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2024, pp. 121–128, 2024.

nitive impairment (MCI) before advancing to full dementia [3-4]. Early detection of cognitive decline associated with MCI and AD is critical, as it can help reduce healthcare costs and lessen the emotional impact on both individuals with AD and MCI and caregivers [5]. While current diagnostic techniques often rely on invasive procedures, such as magnetic resonance imaging (MRI) or cerebrospinal fluid (CSF) analysis, these methods are costly and resource-intensive.. However, AD is associated with specific impairments (alternations) in continuous speech, non-invasive speech and text analysis via artificial intelligence offers a promising and cost-effective diagnostic tool. From a clinical perspective AD is typically characterized not only by structural changes in the brain and the presence of certain proteins but also by alterations in coherent (spoken) speech [6]. Speech in people with AD is often described as "flat," exhibiting reduced lexical diversity, simplified syntax, and frequent repetition [7-8]. Since natural aging also affects speech, it is essential to distinguish age-related language changes from those indicating neurodegenerative diseases , taking into account demographic factors such as age, gender, and education [9-11]. T his study aims to identify lexical predictors for detecting AD and MCI by analyzing data from 214 participants with AD, MCI, or healthy people (CN) status through a picture description task. Linguistic features we extracted from participants´ speech samples and analyzed to compare patterns across these diagnostic groups. Following a review of related work, we present the methodology, data analysis, and results, concluding with interpretations and future implications.

2 Related Work

Numerous studies have identified specific linguistic markers that can effectively distinguish between healthy elder people and individuals with MCI or AD [8, 12-15]. One important linguistic marker associated with AD is lexical diversity, which tends to be lower among individuals with the disease, often resulting in reduced ability to communicate effectively [16]. Lexical diversity is defined as the ratio of unique words to the total number of words in a text or speech sample, with a higher ratio indicating greater range of vocabulary [17].

In addition to lexical diversity, lexical density is frequently examined to further understand language characteristics. Lexical density is based on vocabulary, divided into two categories – lexical (or content) words and function (orgrammatical) words. Lexical words encompass adjectives, adverbs, nouns and verbs, as they primarly convey the meaning of the text. In contrast, function words, such as pronouns, conjunctions, and prepositions, which serve a grammatical purpose rather than contributing to the primary meaning [18]. The author [15] conducted a detailed analysis of the impact of AD on selected lexical features representing various aspects of the lexicon, such as diversity, density, and sophistication (Type Token Ratio, Uber Index, Entropy, etc.). The author aimed to identify the optimal set of features and the most effective input length to maximize the classification model´s accuracy. The findings revealed that, although diversity-related metrics constituted approximately half of the

top 50 features and density-related and specificity metrics despite making up less than 10%, demonstrated superior performance, yielding higher F1 scores than diversity. Authors [19] explored the linguistic characteristics associated with AD by examining differences in the usage of lexical words (nouns, verbs, and pronouns) between individuals with AD and healthy people (cognitively normal, CN) . Their results indicated that individuals with AD used significantly fewer substantive changes than healthy (CN) people ($p < 0.01$). Additionally, they observed, that people with AD exhibited reduced lexical diversity then CN, showing an increased reliance on pronouns and diminished diversity in noun usage.

The above mention studies suggest that lexical markers may serve as effective indicators of neurodegenerative disease symptoms. Although this topic has received substantial attention in recent years [4-8], none of them has focused on an inflected language, such as Slovak.

3 Methodology

A common research method for assessing language deficits in individuals with AD or MCI involves the use of picture description tasks that feature multiple topics [4, 8, 13, 20]. In these studies, researchers applied natural language processing (NLP) techniques to extract specific linguistic features from the speech of both healthy (CN). individuals and those suffering from AD and MCI By identifying differences in these features, they trained a model - a classifier - to categorize individuals as either healthy (CN) or people with neurogenerative diseases. To control for variations in speech that may arise from factors other than disease, researchers typically analyze speech outputs using balanced dataset based on three demographic characteristics (age, gender and education) [21-23]. Considering these factors, we selected transcriptions from picture description tasks completed by both healthy participants (CN) and individuals with early or progressing neurodegenerative disease for our research.

3.1 Participants

The data for this study was sourced from the Early Warning of Alzheimer (EWA) project [24]. Speech samples, recorded through a picture description task, included both healthy individuals and patients diagnosed with cognitive impairments. Participants met criteria such as being over age 50 and free from serious psychiatric or neurological conditions, speech disorders, or severe visual impairment. Healthy individuals scored 24 or higher on the MoCA cognitive function test, while patient participants had a score of at least 18.

Voice recordings were transcribed using an Automatic Speech Recognition (ASR) tool developed by the Slovak Academy of Sciences, followed by manual review. This study analyzes responses from 107 cognitively healthy individuals, 63 with Alzheimer's disease, and 44 with MCI, representing a balanced subset from the larger EWA dataset of around 1,614 participants.

3.2 Metrics

We used a suitable tool from one of the libraries of the Python programming language for the texts obtained from the language expressions of the probands of the individual determined groups. It was a Stanza library for tokenization, lemmatization, and other natural language processing tasks to apply text complexity measures. Subsequently, we automatically extracted 10 linguistic signs (characteristics) from the texts, focusing mainly on lexical density: *V1/W* (*Words that occur only once / Total Words*), *V2/W* (*Words that occur twice / Total number of words*), *H1* (*Hapax diversity measure*), *R1* (*Lexical richness measure, with hapax legomena*), *R2* (*Lexical concentracion measure, with repeated words*), *NOUN/W* (*Nouns / Total Words*), *ADJ/W* (*Adjectives / Total Words*), *ADV/W* (*Adverbs / Total Words*), *VERB/W* (*Verbs / Total Words*).

Metrics *H1, H2, R1* and *R2* were calculated using formullas (1-4):

$$H1 = \frac{\log(T) \cdot 100}{1 - V1/T}, \tag{1}$$

$$H2 = \frac{\log(T) \cdot 100}{1 - V2/T}, \tag{2}$$

$$R1 = \frac{\log(N) \cdot 100}{1 - V1/N}, \tag{3}$$

$$R2 = \frac{\log(N) \cdot 100}{1 - V2/N}, \tag{4}$$

where T – number of unique words, $V1$ – number of words that occur only once, $V2$ – number of words that occur twice, N – number of total words.

3.3 Research Assumptions

The presented article traces the relationship between individual measures of lexical density determined from transcriptions of audio recordings of speech during the picture description task and the level of cognitive impairment of the patients (dgn), while this level was marked by three categorical values, namely (0 - CN,healthy patient, 1 - MCI, and 2 - AD). As part of the experiment, the following assumptions were made:

- Patients with mild cognitive impairment describe situations with lower lexical density (measured by metrics *V1/W, V2/W, H1, H2, R1, R2, NOUN/W, ADJ/W, ADV/W, VERB/W*) compared to healthy patients due to mild difficulties with expressing and processing information.
- Patients with Alzheimer's disease describe a situation with even lower lexical density (measured by metrics *V1/W, V2/W, H1, H2, R1, R2, NOUN/W, ADJ/W, ADV/W, VERB/W*) compared to patients with mild cognitive impairment due to more serious language deficits and cognitive limitations.

4 Results

The null hypothesis (H0) can be formulated as follows: There is no statistically significant relationship between the selected indicators of lexical density and the degree of cognitive disability of the patient.

Due to non-normal data distribution, as indicated by the Shapiro-Wilk test ($p < 0.05$), non-parametric methods were used to assess associations between lexical density measures (*V1/W, V2/W, H1, H2, R1, R2, NOUN/W, ADJ/ W, ADV/W, VERB/W*) and the degree of cognitive impairment of the patient, non-parametric procedures were used.

Table 1 presents the degree of association between the above-mentioned measures and the degree of cognitive level of the respondent (*dgn*) in the observed 214 samples. Significant relationships were found for V*1/W, R2, ADJ/W* ($p < 0.001$), as well as for *H1, ADV/W* ($p < 0.01$).

Table 1: Gamma coefficients: lexical density x dgn

Metric	Valid N	Gamma	Z-score	p-value
V1/W & dgn	214	0.203221***	3.48637	0.000490
V2/W & dgn	214	0.053228	0.91216	0.361683
H1 & dgn	214	0.145262**	2.49158	0.012718
H2 & dgn	214	-0.028523	-0.48882	0.624968
R1 & dgn	214	-0.064127	-1.10081	0.270980
R2 & dgn	214	-0.555665***	-9.53960	0.000000
NOUN/W & dgn	214	-0.014862	-0.25463	0.799012
ADJ/W & dgn	214	-0.278981***	-4.78137	0.000002
ADV/W & dgn	214	-0.155380**	-2.66690	0.007655
VERB/W & dgn	214	0.074857	1.28342	0.199345

Note: *** - $p < 0.001$, ** - $p < 0.01$, * - $p < 0.05$

For further investigation, the assumption of homogeneity of variances was verified by Levene's test. Due to the inequality of variances ($F > 1.783, p < 0.05$), non-parametric tests were subsequently chosen for the comparison of several independent samples.

From Table 2, it can be seen that *R2* can distinguish between healthy people and people with neurodegenerative disease but cannot distinguish a specific disease. Conversely, *V1/W* and *H1* metrics can distinguish healthy patients from those with mild cognitive impairment. The ability to discriminate between healthy patients and patients with Alzheimer's disease was demonstrated by the proportion of adjectives to all words (*ADJ/W*) as well as the proportion of adverbs to all words (*ADV/W*).

Table 2: Multiple comparisons of the measure of lexical density x dgn

Metric	Valid	Sum of Ranks	Mean Rank	0	1	2
V1/W: Kruskal-Wallis test: $H(2, N = 214) = 16.67294$, $p = 0.0002$						
0	108	9915.5	91.8102		**0.000272**	0.053241
1	44	5947.0	135.1591	**0.000272**		0.306096
2	62	7142.5	115.2016	0.053241	0.306096	
H1: Kruskal-Wallis test: $H(2, N = 214) = 11.92536$, $p = 0.0026$						
0	108	10293.5	95.3102		**0.001852**	0.372791
1	44	5862.0	133.2273	**0.001852**		0.186970
2	62	6849.5	110.4758	0.372791	0.186970	
R2: Kruskal-Wallis test: $H(2, N = 214) = 68.12768$, $p = 0.0000$						
0	108	15221.0	140.9352		**0.000007**	**0.000000**
1	44	3901.5	88.6705	**0.000007**		0.098482
2	62	3882.5	62.6210	**0.000000**	0.098482	
ADV/W: Kruskal-Wallis test: $H(2, N = 214) = 7.080144$, $p = 0.0290$						
0	108	12407.5	114.8843		1.000000	**0.033711**
1	44	5025.5	114.2159	1.000000		0.138280
2	62	5572.0	89.8710	**0.033711**	0.138280	
ADJ/W: Kruskal-Wallis test: $H(2, N = 214) = 16.71594$, $p = 0.0002$						
0	108	13370.5	123.8009		0.093404	**0.000202**
1	44	4397.0	99.9318	0.093404		0.616234
2	62	5237.5	84.4758	**0.000202**	0.616234	

5 Discussion

Our study suggests that speech lexical density serves as a significant indicator of patient cognitive health. Specific measures of lexical density, such as the proportion of words that occur only once ($V1/W$), the lexical density index based on unique words ($H1$), the repetition index ($R2$), the proportion of adjectives (ADJ/W) and adverbs (ADV/W), proved to be statistically significant factors associated with cognitive impairment. These findings support the hypothesis that alterations in language production, especially regarding lexical density and part-of-speech selection, are associated with cognitive decline severity and they align with existing research that has identified specific linguistic features associated with neurodegenerative diseases, particularly AD and MCI [8,12-15, 18].

The significance of $V1/W$ and$R2$ in identifying cognitive impairment supports the findings by [4], who noted that increased reliance on basic vocabulary reflect cognitive deterioration in AD patients, particularly in narrative speech tasks.

The findings for $H1$ and ADV/W also align with research [1] indicating that individuals with MCI display lower lexical density than healthy individuals, yet higher than those with AD. This emphasizes the potential of non-invasive, speech-based assessments in early diagnostics across languages.

While Kurdi's study [14] found the ratios of adjectives, adverbs, and verbs to total words significant, our results diverged by showing noun-verb ratios

as weaker predictors of AD. Our findings do align, however, with previous research [25] that showed only adjective and adverb ratios relative to unique words significantly differed among groups.

6 Conclusion

This experiment yielded valuable insights into the relationship between speech lexical density and cognitive impairment, demonstrating that lexical density metrics correlate with varying degrees of cognitive deficit. The presented study is non-invasive, which means that the assessment of language skills takes place without the need to interfere with the patients' physical health. This approach minimizes the psychological stress of testing, allowing patients to express themselves naturally and enhancing data validity. Considering the progressive nature of cognitive diseases such as Alzheimer's disease, our findings could be the basis for the development of new diagnostic methods and interventions that could help early detection and monitoring of cognitive changes. Future research could also explore other aspects of speech complexity for a holistic understanding of language and thought depending on the level of cognitive health.

Acknowledgements. This work was supported by the University Grants Agency (UGA) no. VII/4/2024.

References

1. Williams, E., McAuliffe, M., Theys, C.: Language changes in Alzheimer's disease: A systematic review of verb processing. Brain Lang. 223, 105041 (2021). https://doi.org/10.1016/j.bandl.2021.105041
2. The World Health Organization: Dementia, https://www.who.int/news-room/fact-sheets/detail/dementia (2023, accessed 7 June 2024).
3. Nussbaum, R.L., Ellis, C.E.: Alzheimer's Disease and Parkinson's Disease. N. Engl. J. Med. 348, 1356–1364 (2003). https://doi.org/10.1056/NEJM2003ra020003
4. Fraser, K.C., Fors, K.L., Kokkinakis, D.: Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. Comput. Speech Lang. 53, 121–139 (2019). https://doi.org/10.1016/j.csl.2018.07.005
5. Robin, J., Xu, M., Kaufman, L.D., Simpson, W.: Using Digital Speech Assessments to Detect Early Signs of Cognitive Impairment. Front. Digit. Health. 3, 749758 (2021). https://doi.org/10.3389/fdgth.2021.749758
6. Bose, A., Ahmed, S., Cheng, Y., et al.: Connected speech features in non-English speakers with Alzheimer's disease: protocol for scoping review. Syst. Rev. 13(1), 40 (2024). https://doi.org/10.1186/s13643-023-02379-y
7. Rentoumi, V., Paliouras, G., Danasi, E., et al.: Automatic detection of linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: A computational linguistics analysis. In: 8th IEEE Int. Conf. Cogn. Infocommunications (CogInfoCom). IEEE, pp. 33–38 (2017).

8. Garcia, D.L., Gollan, T.H.: 15 Different Languages, Different Linguistic Markers: Predicting Which Bilinguals will Develop Alzheimer's Disease with Spontaneous Spoken Language. J. Int. Neuropsychol. Soc. 29(s1), 226–227 (2023). https://doi.org/10.1017/S135561772300334X
9. Martínez-Nicolás, I., Llorente, T.E., Ivanova, O., Martínez-Sánchez, F., Meilán, J.J.G.: Many Changes in Speech through Aging Are Actually a Consequence of Cognitive Changes. Int. J. Environ. Res. Public Health. 19(4), 2137 (2022). https://doi.org/10.3390/ijerph19042137
10. Larsson, S.C., Traylor, M., Malik, R., et al.: Modifiable pathways in Alzheimer's disease: Mendelian randomisation analysis. BMJ 359, j5375 (2017).
11. Stern, Y.: Cognitive reserve in ageing and Alzheimer's disease. Lancet Neurol. 11(11), 1006–1012 (2012). https://doi.org/10.1016/S1474-4422(12)70191-6
12. Calzà, L., Gagliardi, G., Rossini Favretti, R., Tamburini, F.: Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. Comput. Speech Lang. (2021). https://doi.org/10.1016/j.csl.2020.101113
13. Lindsay, H., Tröger, J., König, A.: Language Impairment in Alzheimer's Disease—Robust and Explainable Evidence for AD-Related Deterioration of Spontaneous Speech Through Multilingual Machine Learning. Front. Aging Neurosci. 13, 642033 (2021). https://doi.org/10.3389/fnagi.2021.642033
14. Kurdi, M.Z.: Automatic Identification of Alzheimer's Disease using Lexical Features extracted from Language Samples. arXiv preprint arXiv:2307.08070 (2023).
15. Kurdi, M.Z.: Automatic diagnosis of Alzheimer's disease using lexical features extracted from language samples. J. Med. Artif. Intell. 7, 13 (2024). https://doi.org/10.21037/jmai-23-104
16. Fraser, K.C., Meltzer, J.A., Rudzicz, F.: Linguistic Features Identify Alzheimer's Disease in Narrative Speech. J. Alzheimers Dis. 49(2), 407–422 (2016). https://doi.org/10.3233/JAD-150520
17. Baese-Berk, M.M., Drake, S., Foster, K., Lee, D., Staggs, C., Wright, J.M.: Lexical Diversity, Lexical Sophistication, and Predictability for Speech in Multiple Listening Conditions. Front. Psychol. 12, 661415 (2021).
18. Lissón, P., Ballier, N.: Investigating lexical progression through lexical diversity metrics in a corpus of French L3. Discours. 23 (2018).
19. Williams, E., Theys, C., McAuliffe, M.: Lexical-semantic properties of verbs and nouns used in conversation by people with Alzheimer's disease. PLoS ONE 18(8), e0288556 (2023). https://doi.org/10.1371/journal.pone.0288556
20. Cummings, L.: Describing the cookie theft picture: sources of breakdown in Alzheimer's dementia. Pragmat. Soc. 10, 153–176 (2019).
21. Carter, K.C.: Language and Cognition in Mild Alzheimer's Disease. Electron. Theses Diss. 3434 (2022). https://digitalcommons.memphis.edu/etd/3434
22. Sanz, C., et al.: Automated text-level semantic markers of Alzheimer's disease. Alzheimers Dement. Diagn. Assess. Dis. 14, e12276 (2022).
23. Eyigoz, E., Mathur, S., Santamaria, M., Cecchi, G., Naylor, M.: Linguistic markers predict onset of Alzheimer's disease. EClinicalMedicine 28, 100583 (2020).
24. Rusko, M., Sabo, R., Trnka, M., et al.: EWA-DB, Slovak Database of Speech Affected by Neurodegenerative Diseases (2023). https://doi.org/10.1101/2023.10.13.23296810
25. Casnochova Zozuk, N., Munkova, D., Kelebercova, L., Munk, M.: Relationship between language features extracted through NLP and clinically diagnosed Alzheimer's disease and mild cognitive impairment in Slovak (in press).