sciendo
DOI 10.2478/jazcas-2023-0047

LEXICAL DIVERSITY AND LANGUAGE IMPAIRMENT

NATALIIA ČASNOCHOVÁ ZOZUK
Department of Informatics, Faculty of Natural Sciences and Informatics,
Constantine the Philosopher University in Nitra, Nitra, Slovakia

ČASNOCHOVÁ ZOZUK, Nataliia: Lexical Diversity and Language Impairment. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 301 – 309.

Abstract: The development of artificial intelligence tools has seen an enormous growth recently. Linguistic artificial intelligence tools are being successfully applied in the field of speech analysis and discourse. In our study, we used automatic NLP tools to detect differences in picture description in the discourse of people diagnosed with Alzheimer's disease (AD), Mild Cognitive Impairment (MCI) and healthy people. A measure of lexical diversity was used to compare discourse complexity. Transcripts of recordings of the probands within the EWA project were used in the study. From the multiple comparisons, we found that there is a statistically significant difference between healthy people and people suffering from MCI and AD. Our results indicate that healthy people have more lexical diversity than people suffering from MCI and AD – a more diverse vocabulary in spontaneous speech, in our case, when describing a picture.

Keywords: natural language processing, Alzheimer's disease, spontaneous speech, picture description, lexical diversity

1 INTRODUCTION

Nowadays, there is an increasing incidence of civilizational diseases in society affecting the activity of the brain and its cognitive functions, including language and speech. These diseases are included under the collective name – neurodegenerative diseases (Buckner 2004). One of the most serious is Alzheimer's disease (AD). The sooner the diagnosis is made, the sooner methods and means (medicines, therapies) can be applied, which can slow down or even stop further worsening of the condition (Klimova et al. 2015). To determine the correct diagnosis, financially demanding invasive methods are often used, such as MRI examinations or cerebrospinal fluid punctures. However, the symptoms of neurodegenerative diseases are also manifested in the manner and quality of speech of the person impaired, which can be detected by non-invasive methods.

Speech impairment in Alzheimer's disease primarily occurs as a result of a decline in the semantic and pragmatic level of language processing (Ferris – Farlow 2013). Based on the decline in the level of language processing, several language-oriented research methods have been developed to assess the language deficits of people with AD.

One of these methods is the description of a picture that contains several topics (Mueller et al. 2018). A person describes not only the objects and activities, but also emotional states and social ties. Tasks such as describing pictures are often used in scientific research (Lindsay-Troeger – Koenig 2021, Szatloczki et al. 2015). Lindsay et al. (2021) used natural language processing (NLP) methods to extract specific semantic, syntactic, and other linguistic features in healthy people and people with AD, and based on the difference in parameters (language features), they trained a model that classified healthy and impaired people with AD. Frase et al. (2016) used linguistic features to identify AD in narrative speech. They showed some accuracy in the automatic identification of Alzheimer's disease from short speech discourses that were created during the picture description task and revealed significant linguistic features of the speech of healthy and impaired individuals. Jarrold et al. (2014) evaluated the ability of a trained classifier to diagnose dementia subtypes based on spontaneous speech. The findings of Ahmed et al. (2013) indicate that the level of lexical and semantic content and syntactic complexity of the language and speech best describe or reveal the degree of language impairment.

The aim of our study is to compare healthy people and people suffering from MCI and AD based on the lexical diversity of their spontaneous speech discourse when describing a picture.

The study is divided as follows: the following subsections briefly describe the state of the art of the examined issue in Slovakia and define the concept of lexical diversity. The second section is devoted to the research itself, in which we describe methods and procedures. In the third section, we present the results. The research findings are interpreted within the discussion and the final section contains the conclusion of the study.

1.1 Alzheimer's disease research in Slovakia

The EWA[1] (Early Warning of Alzheimer's) research project has been implemented in Slovakia since 2020, the aim of which is to develop a mobile application that would be able, from a person's speech, to detect the presence of early AD symptoms and other neurodegenerative diseases such as Mild Cognitive Impairment (MCI), Parkinson's disease and others. The MCI is the first stage of an incipient neurodegenerative disease, where roughly 25% of cases transform into AD within 5 years. The symptoms of MCI distinguish healthy people from people who evince symptoms of some cognitive problems.

In the EWA project, two types of tasks involving the description of pictures are used to record human speech. In the first type of task, the focus is on appellation of objects or activities that are shown in the picture displayed on a mobile phone. A person has to name what she/he sees using one single word. In the second task, the focus is on a more complex picture description, i.e., the picture contains more

[1] https://www.projektewa.sk/

persons, activities, objects, and relationships. Her/his task is to describe the whole scene of the picture in as much detail as possible. As part of the project, the participants described 65 different images, while over a thousand healthy and over two hundred diagnosed people were recorded, and several tens of thousands of recordings were obtained. However, in our study we will focus only on one more complex picture and its description by a smaller sample of participants.

1.2 Lexical diversity

When cognitive functions decline, language expression or speech discourse is simplified to so-called flat speech, in which linguistic complexity decreases. Complexity is a basic characteristic of a text, depending on many qualitative and quantitative parameters. The latter are the subject of NLP research, as we can determine and quantify them using automatic NLP tools. Within language complexity, we recognize grammatical and lexical complexity. One of the measures for assessing lexical complexity is called lexical diversity which is the subject of our study. From the beginning, lexical diversity (LD) was defined as a ratio of the type and token of the words TTR (Type Token Ratio) (Templin 1957; Johnson 1944), i.e., the total number of unique words (types) is divided by the total number of words (tokens). The closer this ratio is to 1, the greater the lexical diversity of the text. Basically, lexical diversity is the range of unique words used in a text or in speech relative to the overall range of the words in the given text or speech. A larger range corresponds to a higher diversity (Baese – Berk 2021; Durán 2004). This measure is also used as a measure of second language proficiency (Cumming et al. 2005) or vocabulary knowledge (Zareva et al. 2005; Yu 2010), but also as a warning signal or sign of the onset of Alzheimer‘s disease (Garrard et al. 2005; van Velzen – Garrard 2008).

Lexical diversity is calculated according to the following formula:

$$TTR = V/N,$$

where V is the number of unique words and N is the number of all words.

This measure has been proven to be the most suitable for the purposes of our research, because TTR is the most used index of the lexical diversity of a text (Hardie – McEnery 2006).

2 METHODOLOGY

We have no knowledge that similar research has been conducted in Slovakia, expect those mentioned previously. There exists neither research, nor study focusing on lexical complexity as an indicator for detecting neurodegenerative diseases. Therefore, it was necessary to determine which linguistic features (parameters) of language utterance will be investigated and also to define or select participants from the EWA project.

2.1 Participants and materials

For our research, we used the database of texts obtained in the EWA project. Although, Alzheimer‘s disease manifests itself mainly in the elderly population, in the

EWA project, the age of 50+ was chosen as an inclusion criterion. This is due to the fact that the project's task is to investigate early symptoms of the disease, which begin to manifest themselves even at a younger age. We divided our participants into three groups– diagnosed AD people, diagnosed MCI people, and healthy people. People diagnosed with AD and MCI were recruited for the project from specialised medical facilities. Healthy people were recruited through advertising media, magazine advertisements or retirement homes. The participants were informed about the purpose of the project and agreed to provide personal data and speech recordings for scientific purposes.

The inclusion criterion for demonstrating a cognitively healthy mind was the achievement of a specified score in the Montreal cognitive assessment (MoCA) test. Due to the correlation of the occurrence of AD with older age, the average age in the AD group was up to 78 years, while in the group of healthy persons it was only 65 years. A decline in cognitive functions is a natural accompanying phenomenon of human ageing. In order to assess the symptoms of the disease independently of age, balanced groups with approximately the same age means were created. As a result, we included 44 people in the AD group, 57 people in the MCI group, and 204 people in the healthy group.

2.2 Instrument

We used a specific suitable tool from one of the libraries of the Python programming language for the texts obtained from the probands' spontaneous speeches. It was the Natural Language Toolkit (NLTK) library for tokenization, lemmatization, and other tasks related to natural language processing. The lemmatizer developed by LINDAT/CLARIN (the Czech national node of the pan-European research infrastructure CLARIN) with the slovak-snk-ud model was applied from this library. Statistical methods were applied to the obtained values to determine the significance of the differences found.

3 RESULTS

Based on the Mean as well as the Mean Rank (Tab. 1), the differences in lexical diversity between healthy people and people suffering from MCI and AD are visible below.

Diagnosis	N	LD Mean	LD Std.Dev.	LD Std.Err	LD -95,00%	LD 95,00%	LD Sum of Ranks	LD Mean Rank
AD	44	0.75	0.11	0.02	0.72	0.78	29742.00	675.95
MCI	57	0.71	0.11	0.01	0.68	0.73	32523.50	570.59
Healthy person	204	0.66	0.10	0.00	0.65	0.66	370649.50	447.10

Tab. 1. Lexical diversity – mean

In the case of the AD and healthy groups (Tab. 2), we identified significant deviations from normality based on the Shapiro-Wilk W test.

Diagnosis	N	W	p
AD	44	0.98	0.50
MCI	57	0.98	0.49
Healthy person	204	0.92	0.00

Tab. 2. Shapiro-Wilk W test – results

Due to deviations from normality, we will use the non-parametric Levene test (for homogeneity of variances) to test the equality of variances. We reject the null hypothesis of equality of variances stating that there is no statistically significant difference in the variances of the lexical diversity between the three examined groups (Tab. 3).

	MS Effect	MS Error	F	p
Lexical density	0.01101	0.00249	4.42221	0.01226

Tab. 3. Levene test – results

Due to the violation of the assumptions of normality and equality of variances, we use the Kruskal-Wallis test to test the global null hypothesis. Based on the results ($H(2, N = 930) = 39.622$, $p = 0.0000$) we reject the global null hypothesis at the significance level of 0.001, which claims that there is no statistically significant difference between the groups in lexical diversity. After rejecting the global hypothesis, we were interested in groups between which there exists a statistically significant difference. From the multiple comparisons (Multiple comparisons of mean ranks for all groups) we identified two homogeneous groups (MCI, AD) and (Healthy persons) as well as statistically significant differences between healthy persons and persons suffering from MCI and AD (Tab. 4).

Diagnosis	LD Mean	LD Mean Rank	1	2
Healthy p.	0.65808	447.10		****
MCI	0.70575	570.59	****	
AD	0.75083	675.95	****	
Note: **** - Homogenous Groups, $p > 0.05$				

Tab. 4. Multiple comparisons – results

4 DISCUSSION

Although speech impairment is a secondary symptom of AD, many studies (e.g. Bucks et al. 2000; Kavé – Goral 2016; Kavé – Goral 2018) have shown that the

decline in language skills occurs relatively early in people diagnosed with Alzheimer's disease and can serve as a sensitive indicator of the gravity and progression of the disease over time.

It has been shown that the level of lexical diversity is statistically significant for assessing the health of a person's cognitive abilities. In accordance with Kavé and Goral (2018), we also confirmed that the ratio of type and token, in our case, unique and all words, is significantly influenced by the total number of words in utterance. Previous studies (e.g. Bucks et al. 2000; Kavé – Goral 2016) have found that, in general, lexical diversity is lower within the utterance of AD sufferers than healthy people. However, this phenomenon was not specifically confirmed in our study. We believe it is caused by the diagnosed persons describing the picture very briefly. The average number of words used by people diagnosed with MCI was approximately 95 words, compared to only 50 words for those diagnosed with AD. Healthy people used an average of around 120 words, which is a statistically significant difference compared to AD people. It resulted in the finding that the lexical diversity of people diagnosed with AD or MCI is higher compared to healthy people. When using fewer words, the ratio of unique words to all words increases, pointing out that a higher value of lexical diversity in our case does not mean a more complex and rich expression. Here is an example of a picture and the transcription of discourse of the probands of each examined group (AD, MCI, and healthy people).

AD: *"no neviem prečo tam do toho klepe či búcha do toho svetla tam a ach je chlapček zase berie si z oného banán ale sa mu šmykla asi stolička neviem či nespadne tam tam je ešte nejaké.."* ***(37 slov zo 64 slov)***

[AD: 'Well, I don't know why he's knocking or banging on that light over there, and oh, there's the boy again, he's taking that banana from the other one, but maybe his chair slipped, I don't know if he will fall, there there's another one there...' (37 words out of 64 words)]

MCI: *"no v kuchyni decko nejaké tam niečo pustil vodu voda do drezu vyteká z drezu voda vonku vidím tu ďalšie na kraji ešte mačku nejakú mačičku a dotyčný pán rozbil bola buchol do svetla varechou a zase na kuchynskom pulte tam je nejaké nejaký hrniec tiež niečo vyteká vonku nejaká omáčka alebo také niečo..."* ***(54 slov zo 106 slov)***

[MCI: 'Well, in the kitchen, a child has poured water into the sink, water is flowing out of the sink, outside, I see another cat on the side, and the man in question broke it, hit the light with a cooking pot, and there is a pot on the kitchen counter, something is also leaking outside, some kind of sauce or something like that...' (54 words out of 106 words)]

Healthy person: *"chlapec stojí na stoličke naťahuje sa za banánom stolička sa mu prevracia asi padne z vodovodu tečie voda do umývadla vyteká von pozerá sa tam kocúr na to zboku otec má v ruke varechu zdvihol ju chcel trafiť muchu ale rozbil lampu ktorá je visiaca majú tam dve dve police jedna je otvorená polovica dverí sú tam priečky medzitým tam je fľaška ktorá…"* ***(63 slov zo 138 slov)***

[Healthy person: "a boy is standing on a chair, he is reaching for a banana, his chair is tipping over, he is about to fall from the water tap, water is flowing into the sink, it is flowing out, there is a cat looking at it from the side, the father has a cooking pot in his hand, he raised it, he wanted to hit a fly, but he broke the lamp that is hanging. there are two two shelves one half of the door is open there are partitions meanwhile there is a bottle which..." (63 words out of 138 words)]

5 CONCLUSION

Investigating the complexity of human speech may benefit the automatic detection of Alzheimer's Disease symptoms through speech pattern analysis. Differences at the lexical level between the speech of a person diagnosed with AD and the speech of a healthy person can be captured and quantified. However, it is necessary to know which lexical parameter is suitable for a specific task of speech analysis. It was evident that the lexical diversity of AD or MCI people is higher for a short speech utterance describing a picture, which, however, does not represent the richness of the speech utterance. This is an interesting and scientifically significant finding.

ACKNOWLEDGEMENTS

During the research work, we were able to use the data obtained within the EWA project, for which we are very grateful to the researchers of this project.

References

Ahmed, S., Haigh, A. M., Jager de, C. A., and Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. Brain: a journal of neurology, 136(12), pages 3727–3737.

Baese-Berk, M. M., Drake, S., Foster, K., Lee, D., Staggs, C., and Wright, J. M. (2021). Lexical Diversity, Lexical Sophistication, and Predictability for Speech in Multiple Listening Conditions. Front. Psychol. Vol. 12. Accesibble at: https://doi.org/10.3389/fpsyg.2021.661415.

Buckner, R. L. (2004). Memory and executive function in aging and AD: multiple factors that cause decline and reserve factors that compensate. Neuron, 44, pages 195–208.

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

Covington, M. A., and McFall, J. D. (2010). Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). Journal of Quantitative Linguistics, 17, pages 94–100.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., and Jamse, M. (2005). Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL®. ETS Res. Rep. Ser. 2005, pages 1–77.

Durán, P., Malvern, D., Richards, B., and Chipere, N. (2004). Developmental trends in lexical diversity. Applied Linguistics, 25(2), pages 220–242.

Fergadiotis, G., Wright, H. H., and Green, S. B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. Journal of Speech, Language, and Hearing Research, 58(3), pages 1–13.

Ferris, S., and Farlow, M. (2013). Language impairment in Alzheimer's disease and benefits of acetylcholinesterase inhibitors. Clin. Interv. Aging 8, pages 1007–1014.

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic Features Identify Alzheimer's Disease in Narrative Speech. Journal of Alzheimer's Disease, 49(2), pages 407–422.

Garrard, P., Maloney, L. M., Hodges, J. R., and Patterson, K. (2005). The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. Brain 128, pages 250–260.

Hardie, A., and McEnery, T. (2006). Statistics. In K. Brown (ed.).: Encyclopedia of Language and Linguistics, 2nd edition. Amsterdam: Elsevier, pages 138–146.

Jarrold, W., Peintner, B., Wilkins D., Vergryi D., Richey, C., Gorno-Tempini, M. L., and Ogar, J. (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. Baltimore, Maryland, USA. Association for Computational Linguistics.

Kavé, G., and Dassa, A. (2018). Severity of Alzheimer's disease and language features in picture descriptions. Aphasiology, 32(1), pages 27–40.

Johnson, W. (1944). Studies in language behavior: a program of research. Psychol. Monogr., 56, pages 1–15.

Kavé, G., and Goral, M. (2016). Word retrieval in picture descriptions produced by individuals with Alzheimer's disease. Journal of Clinical and Experimental Neuropsychology, 38(9), pages 958–966.

Klimova, B., Maresova, P., Valis, M., Hort, J., and Kuca, K. (2015). Alzheimer's disease and language impairments: social intervention and medical treatment. Clin. Interv. Aging, 10, pages 1401–1407.

Lindsay, H., Tröger, J., and König, A. (2021). 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. Accesibble at: https://doi.org/10.3389/fnagi.2021.642033.

Mueller, K. D., Hermann, B., Mecollari, J., and Turkstra, L. S. (2018). Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. J. Clin. Exp. Neuropsychol., 40(9), pages 917–939.

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., and Pakaski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. Front. Aging Neurosci., 20(7). Accessible at: https://doi.org/10.3389/fnagi.2015.00195.

Templin, M. C. (1957). Certain Language Skills in Children; Their Development and Interrelationships. Minneapolis, MN. University of Minnesota Press.

Velzen van, M., and Garrard, P. (2008). From hindsight to insight – retrospective analysis of language written by a renowned Alzheimer's patient. Interdiscipl. Sci. Rev., 33, pages 278–286.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. Appl. Linguist., 31, pages 236–259.

Zareva, A., Schwanenflugel, P., and Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: variable sensitivity. Stud. Second Lang. Acquisit., 27, pages 567–595.