Doctoral School of Information and Biomedical Technologies

Polish Academy of Sciences (TIB PAN)

SUBJECT: **Morphosyntactic speech patterns as markers of the early stages of dementia (using corpus analysis and LLM models).**

SUPERVISOR: . Agnieszka Mykowiecka (IPIPAN)

DESCRIPTION:

It is estimated that there are approximately 50 million people worldwide living with dementia. This number is steadily rising, and according to WHO projections, the number of people with dementia will triple by 2050. While we currently cannot prevent dementia, its early detection can be helpful in many ways.

Dementia affects language use, so detecting changes can serve as an early indicator of the developing disease. For this reason, research into these relationships is of interest to many researchers. In the study (Hong Lei, Zhanhao, 2025), a bibliometric analysis was conducted of 545 articles (from 1994–2023) on this topic. The authors noted that most publications originate from the United States, with authors also coming from the United Kingdom, Australia, and Canada, and the most prominent journals being “Aphasiology” and “Brain and Language.” Nine themes were identified, ranging from semantic and syntactic processing to natural language processing (NLP) techniques and speech therapy.

The authors’ analysis indicates that linguists have addressed a variety of issues related to dementia, including semantic analysis, multilingualism and cognitive functions, primary progressive aphasia and speech apraxia, natural language processing techniques, the role of speech-language pathologists, communication dynamics in various contexts, speech processing, syntactic processing, and word retrieval. Many open questions continue to pose a challenge for researchers of English and have so far received virtually no analysis in the context of Polish data. The proposed dissertation topic involves collecting data and conducting research on the linguistic characteristics of text produced by people diagnosed with dementia. Due to the nature of access to medical data, the scope and type of research will depend on the volume and type of collected material. In addition to purely corpus-based research, there are also plans to train models to detect speech patterns in people with the condition. For example, the research may focus on: pauses and hesitations, syntactic simplification, reduced vocabulary, and the more frequent use of pronouns instead of nouns. Among more advanced language features, the research may concern maintaining narrative coherence, conducting dialogue, and describing images. Sample research questions include: Can corpsu analysis data or LLMs detect dementia earlier than traditional

tests? Which language features are truly diagnostic? Do the models perform equally well across different languages (English and Polish)? How can we ensure the explainability of AI decisions?

Candidates should hold a master's degree in computer science, engineering, or philology/linguistics/psychology, and have experience in data processing and the application of machine learning methods, particularly in the area of deep learning frameworks (e.g., PyTorch) and language models. A scientific curiosity and a willingness to learn are essential.

Hong Lei, Zhanhao Jiang, Linguistic insights into dementia from 1994 to 2023: A structural topic modeling-assisted bibliometric analysis, Language and Health, Volume 3, Issue 2, 2025, 100064, ISSN 2949-9038,

Heybe, M., Gibson, L., Price, A.C. *et al*. Identifying people with potentially undiagnosed dementia with Lewy bodies using natural language processing. *npj Aging* **11**, 68 (2025). https://doi.org/10.1038/s41514-025-00252-x

Lotem Peled-Cohen, Roi Reichart; A Systematic Review of NLP for Dementia: Tasks, Datasets, and Opportunities. *Transactions of the Association for Computational Linguistics* 2025; 13 1204–1244. doi: https://doi.org/10.1162/TACL.a.35