PhD Thesis Supervisor's report

Thesis author:	Mgr. Emil Svoboda
Thesis title:	Modelling Compounds for Multilingual Data Resources
Supervisor:	Mgr. Magda Ševčíková, Ph.D.

Thesis contents summary

In Emil Svoboda's PhD thesis, the word-formation process of compounding is approached from multiple perspectives. It interconnects the extensive and multifaceted linguistic debate with the treatment of compounds in language data resources and tools:

- In Chapter 2, a selective survey of accounts of compounding in linguistic literature delineates this word-formation process with respect to bordering phenomena and summarizes the main features employed in existing compound taxonomies. The thesis points out that data and tools lag behind the theoretical considerations. Datasets capturing compounds are limited across languages; computational tools that would process compounds in a linguistically adequate way are essentially non-existent. The goal of the thesis is to work towards filling these gaps.
- Chapter 3 presents Emil's original research in developing a series of tools for automatic processing of compounding. The tools have been gradually expanded in terms of the number of word-formation phenomena covered and of the number of languages. While the first tool (Czech Compound Splitter) parsed a Czech compound into the words from which it was created, the second tool (Word Formation Analyzer for Czech) was capable of distinguishing Czech compounds from primary words and outputs of derivation and of identifying the motivating items both for compounds and derivatives; the third, most advanced tool (Parent Retrieval Neural Tool) eventually extended these functionalities to seven languages.
- The fourths and fifth chapter describe another area of Emil's PhD research, namely the annotation of compounds in two types of linguistic data. First, the annotation scheme is described, the application of which to the DeriNet lexical database led to a coherent representation of standard and neoclassical compounds. The second type of data is syntactically annotated corpora, which Emil identifies as adequate for capturing the internal structure of compounds. For his pilot study, he chooses the widely used multilingual collection Universal Dependencies. He convincingly shows that the annotation scheme could be adapted to capture the basic distinctions discussed in the literature and allow for comparison across languages.

Work progress and evaluation

Emil Svoboda started his PhD studies at the Institute of Formal and Applied Linguistics four years ago after completing his Master's degree in Phonetics at the Faculty of Arts, Charles University. He was able to expand his strong linguistic background by intensively studying language technology issues. It is this solid knowledge in both fields that gave rise to his original goal of designing tools for automatic processing of compounds. To achieve this, Emil worked very independently and with great dedication, often with considerable time commitment to overcome computational and linguistic challenges.

Although the first and second year of his studies were complicated by the pandemic, Emil completed a three-month internship under Prof. Martin Haspelmath's supervision at the Max Planck Institute for Evolutionary Anthropology in Leipzig and another three-month internship at Harvard University in Boston. At the Institute of Formal and

Applied Linguistics, Emil participated as a team member in the research grant project Start and was the Principal Investigator of the Charles University student grant GAUK in 2022 and 2023. For the latter grant, he received the Award of the Chairman of the Charles University Grant Board, which will be presented on this year's International Students' Day in November.

In the course of his PhD studies, Emil published six papers in scientific journals and peer-reviewed conference proceedings and co-authored two datasets, which have been made publicly available through the LINDAT/CLARIAH-CZ repository. One of his papers, entitled *PaReNT (Parent Retrieval Neural Tool): A Deep Dive into Word Formation Across Languages* (co-authored by Magda Ševčíková), received the Outstanding Paper Award at the LREC-COLING 2024 conference; it was one of ten papers awarded this prize out of a total of 1,555 papers presented at the main conference.

Recommendation

Overall, I consider Emil's thesis to be a valuable and original scientific contribution to word-formation research, which stands up to international scrutiny. I recommend it as fully adequate for the PhD defense.

Prague, September 17, 2024

Mgr. Magda Ševčíková, Ph.D.

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University
Malostranské náměstí 25
118 00 Praha 1