Nabil Hathout
Directeur de recherche CNRS
CLLE, Université de Toulouse
Nabil.Hathout@univ-tlse2.fr

Toulouse, August 28, 2024

Doctoral Thesis Review

Thesis: Natural Language Correction With Focus on Czech
Author: Emil Svoboda
University: Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Supervisor: Magda Ševčíková
Reviewer: Nabil Hathout

Emil Svoboda's thesis deals with the analysis and description of morphological compounds in Czech and other languages.

Compounds represent a significant proportion of lexemes in Slavic languages such as Czech and Russian, but also in Sanskrit, Latin and Germanic languages such as English and German. While English and German language resources such as CELEX and GermaNet provide extensive descriptions of compounds, similar resources do not exist for Czech (and Russian). Emil Svoboda's thesis makes a number of contributions to help reduce this gap, with a particular focus on the Czech language. To this end, Emil Svoboda proposes a set of tools for identifying compounds and their components which he calls parents. He has also added around 4,000 fully analyzed compounds to the DeriNet resource, tripling the number of fully analyzed compounds in the resource. The task addressed by Emil Svoboda is particularly complex, because it presupposes the ability to recognize compounds, to identify their components and then to produce an expression that glosses the meaning of the compound. In addition, compounds can have derivatives, and derivatives of compounds are derivatives, not compounds.

There is no clear, consensual definition of what a compound is. It's not always easy to tell the difference between a derivative and a compound (is *milkman* a compound or a derivative). Compounds are not always simple concatenations of their components. They can contain linking elements. This makes compound analysis a difficult task, especially when performed out of context. In context, it is possible to better grasp the meaning of a compound and the semantic contribution of its components. The difficulty lies in the fact that we don't have formal indices for compounds (they are not signaled by exponents like derivatives), nor do we have clues to identify the semantic relation established between

the components. Without sufficiently precise knowledge of the meaning of the compound and its components, it is difficult to reconstruct a phrase that contains the components and glosses the meaning of the compound.

Emil Svoboda distinguishes several subtypes among compounds. These subtypes exist in Czech and other languages. A first subtype corresponds to neoclassical compounds containing one or more Greek or Latin components. These compounds are part of the learned lexicon, and many of them are borrowings. A second sub-type brings together adverbial compounds, which are produced using rules similar to those used for derivation. These compounds are part of stable formal and semantic patterns, enabling the construction of a large number of compounds.The third subtype is made up of general compounds formed from two or more lexemes. The existence of these subtypes is a further source of complexity, especially as they are not necessarily formally distinguished.

The aim of Emil Svoboda's thesis is twofold. The first is experimental: creation of analyzers and evaluation with respect to datasets. It is in line with current NLP practices, particularly in terms of performance improvement and comparison with the state of the art. The second is resource enhancement based on tools and an annotation guide. The thesis is complemented by a state-of-the-art review of compound analysis, perceptrons, sequence2sequence networks and transformers. The thesis also proposes a syntactic approach to compounds, aimed at refining their description in the corpora that make up the UD resource.

The thesis is based on a body of research that has been published, 2 in journals and 5 in workshop and conference proceedings. It consists of 6 chapters.

Chapter 1 presents the subject and the aims of the thesis: identification of compounds, identification of the components (i.e., parents) they contain, morphological analysis of compound and derived lexemes, the creation of a new version of DeriNet. The introduction also presents the various tools developed and the content of the thesis. Emil Svoboda argues that analysis and processing must be multilingual because composition is a phenomenon found in many different languages, but he does not explain why, for a particular language, a multilingual analysis is superior to a monolingual one.

Chapter 2 presents the framework of the study. It includes a relatively complete presentation of compounding and a definition of the compounds considered in the thesis, which follows in part that of Haspelmath (2002). Section 2.2 deals with compound classification. Section 2.3 presents the resources used to extract the datasets used in the thesis: CELEX2, DeriNet, GermaNet, Golden Compound Analyses, MORBO/COMP, UniMorph, Wiktionary, Word Formation Latin. The presentation could be more systematic, given the diversity of formats and content. For English, a presentation of LaDEC (The Large Database of English Compounds) is probably missing. Moreover, the dataset created as part of the thesis could serve as a benchmark for a shared task of compound analysis. Section 2.4 presents a set of compound parsing tools. Probably missing is a presentation of Marelli's (2023) work on compounds. The section does not indicate which ideas have been taken over from previous work (network architecture, use of word piece embeddings, etc.).

Chapter 3 presents several tools developed by the author for compound analysis. Section 3.2 presents the *Czech Compound Splitter* designed to analyze compounds, and section 3.3 the *Word Formation Analyser for Czech* (WFA.ces) designed to analyze both

derivatives and compounds. Section 3.4 presents one of the main contributions of the thesis, the PaReNT morphological analyzer implemented as a neural network with an architecture designed by Emil Svoboda. PaReNT is a multilingual system for analyzing compounds and derivatives in seven different languages. It features a character-level encoder, BPEmb-based word piece embeddings and a language encoder.

In section 3.4.3, Emil Svoboda analyzes the system's errors and concludes that morphemes are a type of object found in all languages and are relevant to morphological description. This position is not sufficiently well supported, and it cannot be ruled out that what Emil Svoboda is discussing is an artifact of the method. Section 3.4.4 discusses language embedding. Languages are represented using a one-hot encoding, and it seems to me that this encoding takes no account of the other information provided as input to the model. So it seems expected that it doesn't allow us to say anything about the similarity of languages.

Chapter 4 presents version 2.2 of DeriNet, which adds 1115 entries and 4557 annotations (neoclassical components, compounds and derivatives). These modifications have all been manually checked. The new version of DeriNet is one of the contributions of the thesis. These figures are relatively low because the precision of the PaReNT analyzer was prioritized over recall. Accuracy is crucial when the results of analysis are intended to feed a reference resource. However, as the annotation is manually edited, a lower accuracy may be acceptable. Overall, DeriNet contains a little over 45,000 lexemes annotated as compounds. Emil Svoboda estimates that 53,000 DeriNet entries are compounds. While Emil Svoboda's contribution is significant, the annotation of compounds in DeriNet remains partial.

Chapter 5 proposes a new representation of compounds in UD, based in particular on a specialization of the "compound" relation that specifies the dependency that exists between components.

In this thesis, Emil Svoboda presents a body of research on a particularly complex issue: the analysis and representation of compounds. The study confirms the difficulty of the task and makes a significant contribution on several aspects of the question. It also paves the way for future research leading to new advances in compound processing. The doctoral thesis of Emil Svoboda proves his ability to conduct valuable research on difficult questions. It fulfills requirements to seek a Ph.D. degree in computer science.