Review of Thesis Proposal

Reviewer: doc. Mgr. Pavel Štichauer, Ph.D. (ÚRS FF UK)
Thesis title: Modeling Compounding for Multilingual Data Resources
Candidate: Mgr. Emil Svoboda
Supervisor: Mgr. Magda Ševčíková, Ph.D.

The proposal for the PhD thesis, titled *Modeling Compounding for Multilingual Data Resources* submitted by Emil Svoboda, provides an excellent summary of the overall structure of the planned dissertation. It focuses on both the general debate on compounding and the computational tools developed to model compound properties. The project's ambitious goal is clearly formulated at the beginning of the proposal: to develop a multilingual computational tool capable of identifying compounds, segmenting their internal structure, and retrieving the individual constituents (termed "parents") across multiple languages. This broad typological scope, along with a detailed discussion of the relevant data and the identification of obstacles and challenges, makes the thesis proposal a substantial undertaking.

In the following sections, I will provide a brief summary of the proposal's content and highlight some of the intriguing aspects. Due to my limited expertise in the field of computational modelling, I will focus on general aspects leaving the assessment of the technical part to my colleagues on the committee.

Outline the proposal

The proposal is divided into five main sections, which could perfectly function as future chapters of the planned thesis. Indeed, all five sections thoroughly cover the principal topics of the project.

The introduction includes a working definition of compounding, a list of languages to be included in the sample, a significant distinction between a static perspective and a dynamic one. The former perspective is adopted for the project's purpose, given the focus on working with the existing lexicon. Additionally, there is a note on the formal representation of the examples.

The second section is dedicated entirely to presenting various theoretical frameworks that have tackled classifications of compounding. The overview provided here is concise but encompasses all relevant aspects of the issue, including constituent relations, headedness, and endo/exocentricity. The section then delves into well-known challenges, addressing topics such as the boundaries between compounding, derivation, and inflection, morphological variation, as well as two major compounding processes (or patterns): parasynthetic and neoclassical compounding.

The third section aptly reviews the primary data resources and computational tools, including even those that might be considered somewhat outdated (not due to content, but because they are no longer available or freely accessible), such as the interesting project MORBO/COMP.

The fourth section outlines the experimental and computational aspects of the project. In this section, the proposal introduces three main tools, along with their corresponding datasets. These tools are the Czech Compound Splitter, the Word Formation Analyzer for Czech (noteworthy for extending the model to handle all complex words, not just compounds), and PaReNT (acronym for 'Parent Retrieval Neural Tool'). The PaReNT tool is subsequently presented in greater detail, and its performance is evaluated in comparison to

Dummy and ChatGPT. A significant subsection is dedicated to the error analysis of PaReNT. Here, an interesting taxonomy of errors is presented, aptly labelled by Svoboda as data conflict, inflectional confusion, morphological ambiguity, neural hallucination, overretrieval, false morphemes (I would suggest using 'false morphemeness' or something similar to align with the other labels), and semantic irrelevance. These issues are intriguing not due to their unveiling of unexpected behaviour, but rather because they seem to exhibit familiar problems recently addressed by various morphological theories. Notably, these problems align with what has at times been referred to as 'item-and-arrangement' morphology.

The final section is more than a mere summary, since it also introduces a potential future enhancement of the model. Specifically, it addresses the need to incorporate the representation of constituent relationships (i.e., coordination, subordination, and attribution, following the classification proposed by Bisetto & Scalise 2005).

Questions and discussion

I will now proceed with a series of questions that focus on more general aspects. To begin with, I want to acknowledge the clarity with which the proposal has been written. However, I believe there could be opportunities to rephrase certain expressions.

1) For instance, the proposal states at the outset that Czech is chosen as the starting point because "[the author] happen[s] to have the most insight into the language". However, it is worth noting explicitly that this choice is not solely due to the author's native proficiency but also because Czech is fortunately extensively documented and well-suited for the project's objectives.

2) On p. 2, the righthand column at the bottom, the definition of centricity might benefit from greater precision, since "inherit[ing] form and meaning" could be slightly misleading (in line with the literature on this intricate subject, it is understood that 'form' refers to 'formal or grammatical features' which are said to *percolate* onto the entire compound). A similar case can be found on p. 3 (2.2.2), where it is more precise to say that the constituent inflects for number (and the form realizes the values 'plural'). However, I acknowledge that such meticulous precision might not be crucial for the current proposal.

3) I am curious about whether the orthographic issues mentioned on p. 5 are truly irrelevant or not. It appears to me – taking into account my limited expertise – that computational models can only manage sequences that are written as a single integrated word (and when identified as such, the likelihood of them being compounds increases). I would appreciate clarification on this matter.

4) I believe that including additional examples for certain error groups discussed in section 4.3.3 would be beneficial. In particular, I am interested in understanding what could be classified as a "lexeme that is correct but disagrees with the label" (p. 8, Type 1. Data conflict). Likewise, on p. 9, under Type 4: neural hallucination, which, by the way, is a felicitous term for a broad spectrum of baseless formations, could you provide an example of a character switch?

5) Type 5: Overretrieval is also quite interesting. I have another question that might seem computationally naive, but the scenario involving the retrieval of the initial base (that is, the model, as Svoboda puts it, "does not return the parent of the input, but the parent of the parent of the input") is surprising. I would expect that performing such a nested operation could be more complex for these models.

6) Type 7: Semantic irrelevance is also quite intriguing, and my question is once again similar to the one above: if *Fahrzeug* is part of the existing lexicon (alongside its two independently existing bases), why doesn't the model choose this form first (thus adhering to

the higher criterion of what I might term a ‘canonical compound’, which usually consists of only two constituents)?

7) Finally, the future outlook related to dependency structures within compounds appears, in my view, to be complex on theoretical grounds as well. In Distributed Morphology, for instance, roots are not initially assigned a lexical category; the category arises after adding certain functional (and derivational) features. Consequently, I wonder whether the structures presented in Fig. 2 (p. 10) would be unanimously accepted. In particular, I find it difficult to assign NOUN to the root *mal-* which, in turn, can serve as a root for several other related words, such as *malovat*, *malíř, malování*, and so on. The classification of it as a noun with the meaning of “painting” is established by its whole-word form *malba*. The classificatory dependency relation can thus be established between *olej* and *malba*, and not between *olej* and *mal-*. However, it is well possible that my remarks overlook the fact that the structure is always embedded within the entire compound and not isolated, as I initially interpreted it.

Conclusion

In conclusion, I find the PhD thesis proposal submitted by Emil Svoboda to be not only feasible but I also view it as a substantial endeavour that has the potential to provide fresh insights into various intricate aspects of the captivating process of compounding.

In Prague, 28 August 2023

Pavel Štichauer
reviewer of the thesis proposal