Review of thesis proposal

Reviewer: Václav Cvrček, vaclav.cvrcek@ff.cuni.cz
ÚČNK FF UK, Panská 7, Praha

Date: February 6, 2023

Thesis title: Formalisation of the word-formation meaning in language data resources
Candidate: Lukáš Kyjánek
Supervisor: Mgr. Magda Ševčíková, Ph.D.

The thesis proposal submitted by Lukáš Kyjánek is anchored in NLP, specifically language annotation and analysis. The aim of the thesis is to contribute to the analysis and description of word-formation in different languages.

The task is not a new one – there are dozens datasets for languages including Czech that attempt at formalize mostly the derivation (and to some extent other means of word-formation) and also some tools for automatic analysis. However, the situation is still far from optimal, as the datasets usually lack the formal annotation of the word-formation meanings.

Content

The proposal is written in English and consists of seven sections (excluding abstract and references).

In the introductory section Lukáš Kyjánek describes the goal of the thesis as discriminating between word-formation and lexical meaning, formalizing the former one (by harmonizing the existing annotation cross-linguistically) and implementing it on existing datasets modeling word-formation relations between lexemes (e.g. DeriNet). The fact that the author wants to cover all three major word-formation processes (affixation, conversion, and compounding) adds to the ambition of the project.

The second section summarizes approaches to word-formation that were influential esp. in our (Czech) environment. Starting from Apresjan's Lexical functions through the "Czechoslovak" tradition (Dokulil, Daneš, Štekauer), to cross-linguistic approaches of Haspelmath and Bagasheva (comparative semantic concepts). Special attention is paid to approaches dealing with less-commonly studied processes of conversion (2.2) and compounding (2.3).

The next section examines the data sources which are available for the research in word-formation (two for Czech and one for Croatian, French and English). All sources are briefly described – the size of the dataset and the word-formation annotation used (the only exception being Démonette (3.2) where the information about the size is missing).

Empirical experiments related to word-formation are presented in section 4. First, it is the question of granularity of word-formation meanings which is an attempt to delineate the borderline between inflection and derivation. This was explored by comparing the vector

representations of word pairs differing in inflectional or derivational categories. Second experiment pertains to the competition of homonymous affixes for one word-formation meaning; the case study focuses on nominal agent suffixes in Czech and their predictability based on several formal-linguistic features.

Section 5 discusses different approaches on labelling (or the interpretation of the links between related words). The description focuses mostly on machine learning models trained on manually annotated data from DeriNet (Czech). The results are then applied in multi-lingual settings using machine translation techniques, so far with limited success (mainly due to the poor performance of the MT).

This attempt sets the stage for section 6, which discusses comparative cross-linguistic aspects of the research in general. The main task here is to identify formal equivalents for the same word-formation meaning in different languages.

Future perspectives and plans for the dissertation are described in the final section. As far as the formalization itself is concerned, distributional semantics looks – according to the author – as the most promising path. As for the labeling, due to the lack of training data, semi-automatic or unsupervised techniques seem to be the most viable option. Finally, the question of cross-linguistic comparison or language transfer which seems to be intuitively the most complex one is discussed – the current results are unsatisfactory but some suggestions to circumvent these shortcomings are presented.

Evaluation

The research problem (word-formation meaning formalization) is relatively well described in the proposal, however, I would suggest avoiding the diachronic metaphor (new meanings created from old words). Since vast majority of the words under scrutiny exist for a long time in language, I would recommend describing word-formation as formal/semantic *relation* between words.

Questions for discussion:

- All the theoretical frameworks mentioned in the lit. review stem from compositional approach to meaning (e.g. base + mark = new meaning). Have you also considered the discriminative approach (Baayen, Ramscar) and would it be possible to use it for the purpose of annotation?
- The approaches that will serve as the basis for harmonization and the creation of a universal schema for formalizing word-formation meanings are described in Section 2 almost without any evaluation or critical reflection (apart form a note on Bagasheva’s approach in 4.1). What will be the criteria of harmonization, for picking (or not picking) one nomenclature over the other?
- The experiment in 4.1 is conducted on bootstrapped samples of word (lemma) pairs. How was this sampling done? Was the sample drawn from the corpus (text) or a list of types (dictionary)? The former has the ability to represent the distribution in the text (with frequent pairs more likely to be chosen) while the latter gives more

opportunities for lower-frequency items to be selected and thus prefers diversity of the sample over its representativity.
- Results in Figure 1: what I find particularly surprising is the result of negation for adjectives (I assume that "A ~ A" stands for two adjectives in a pair) and verbs. According to this experiment adjectival negation seems to be closer to the inflection pole than the verbal negation which comes out as more word-formation-like category. My experience and research[1] tells me quite the opposite. Do you have an explanation for this?
- The results of automatic labelling described in sec. 5 (as far as the Figure 3 is concerned) looks rather scarce. Maybe it is just a wrong impression based on the figure but most of the links between words in the network remain unlabeled. What is the precision and recall of these methods?
- I did not find interpretation of the results presented in sec. 6 (Table 7 and Figure 4). Some of the similarity metrics reveal some kind of genetic kinship but since the Czech served as a pivot language (from which all other languages were translated) the results might be biased in this respect.

The text is clearly written, only in a few places I find wording or language infelicities:

- In sec. 2.2, p. 4: "For example, the masculine (!) verb *bubnovat* 'to drum' motivated by the noun *buben* 'drum' in Czech can be easily used as a subject of the sentence (*Buben se protrhl*. 'The drum burst.') while the verb can be used as a predicate (*Déšť hlasitě bubnoval na střechu*. 'The rain drummed loudly on the roof.')." First, *bubnovat* is not a masculine verb, secondly, I doubt that it is motivated by *buben* (I would say it is the other way round, if you insist on directionality of the derivation), the wording suggests that *bubnovat* can be used as a subject (which of course it can, but it is not the case of the sentence *Buben se protrhl*) and most importantly, I doubt that *buben-bubnovat* is a case of conversion.
- Sec. 4.2, p. 7: "features that have the potential to a (!) play role"
- Sec 5, p. 9: "The preliminary results of machine translation show that there is translation methods suffer (!) from many different aspects."

Conclusion

It is not clear from the proposal whether the aim of the thesis is to provide an exhaustive "guide" to the formalisation of word-formation meaning or whether it will focus on the analysis and annotation of selected parts of the word-formation system. Secondly, it is not entirely clear whether the goal of the dissertation should be only the design of possible solutions or also their implementation for a specific language.

---

[1] Kováříková, D. - Chlumská, L. - Cvrček, V. (2012): What belongs in a dictionary? The Example of Negation in Czech. Euralex 2012. Oslo. https://euralex.org/publications/what-belongs-in-a-dictionary-the-example-of-negation-in-czech/

Therefore, although I find this a promising project that is very timely in its subject matter and could bring about a number of improvements in the field of NLP, I would consider it appropriate if these basic contours were clear and clearly described in the proposal.

Despite all the above-mentioned objections and unresolved issues, which stem mainly from the fact that the proposal is written in the middle of the dissertation project, I am convinced that Lukáš Kyjánek has demonstrated that he is well oriented in the topic, knows the current approaches and is able to come up with innovative solutions for his dissertation.