Opponent's review of Master thesis

Opponent: Jiří Hana

Thesis: Morphological Segmentation in Czech using Word-Formation Network

Author: Jan Bodnár

The thesis presents a solid, useful and interesting work. It was accepted for publication by one of the top conferences in the field. However, the presentation feels like the author wrote the thesis in a great hurry and did not have time to proofread it: the formatting is very poor and inconsistent, the selection of related work is rather limited and the theoretical background is described in an uneven level of detail.

Despite these significant shortcomings, I do recommend the thesis for defense.

Below are my comments addressing the content of the thesis, its form and the attached code.

Comments about the content

1. Related Work

- This chapter uses a lot of terminology introduced in Chapter 2. It would be better to switch the order of these chapters.
- Bayesian approaches should include:
    - MAP models (mentioned later in 1.1.3 without stating that they fit into the Bayesian paradigm), especially those by Creutz & Lagus (2007), and
    - Snover & Brent (2001).
- I think that the author could have spent more time with some of the most important projects (Goldsmith, Creutz & Lagus), especially providing some assessment.
- Also, there are far more papers dealing with morphological segmentation – see, for example, Table 2 in (Hammarström & Borin 2010); at least some of them should be mentioned.

2. Theoretical background

2.1.1. Czech language

- The possessive suffix *-ův-* is not a good example of inflection because:
    - it is very easy to argue that *-ův-* is a derivational affix,
    - it isn't an ending (instead, it is followed by an ending, say, in *otc-ov-a*), and
    - similar affixes are used by non-inflectional languages.

2.1.2 Morphemes

- I am not sure why the author dives so relatively deep into circumfixes. Other types of affixes are only mentioned or not even that (infixes).
- The author mentions cranberry morphemes, but he should have explicitly stated that they violate the definition of a morpheme from the previous page (analogously, there are also zero morphemes and thematic vowels that violate it as well).

2.1.3 Allomorphy:

- One of the paragraphs describes suppletion, it should be mentioned by name.
    - Yes, *jít* and *šel* are historically two different verbs, and so are *go* and *went.*

2.1.4 Word formation

- Affixation and inflection/derivation are orthogonal processes.
- The distinction between derivation and inflection is important for the author's work. Yet it is only mentioned in one paragraph at the end of the word formation section. More criteria are needed to

distinguish between derivation and inflection than just meaning (see, e.g., Kroeger 2005:253). Some of the examples are kind of problematic (it is not that hard to argue that the difference between *hladký* `smooth' and *hladce* `smoothly' is only grammatical; such examples should be either avoided or accompanied by a comment).

Comments about the format and style

- It is obvious that the author struggles with (La)TeX: bullet lists, quotes, spacing, bibliography, etc. Arrows are printed as *-?*, the output contains boxes marking overflown lines, etc. Why didn't he use MS Word? Or look at some LaTeX tutorials.
- Language examples should be presented in a consistent manner. Czech examples need English glosses. The typical format used in linguistics would show Czech words in italics, followed by an English gloss in single quotes (e.g., *kočka* 'cat'). The author sometimes mixes formats even within a single list of examples.
- Language names are spelled with lower-case in Czech (e.g., *finština*, p. 14).
- It would be nice to provide some general overview at the beginning of chapters, explaining what's coming (the single sentence "We first examine the methods used by different authors." in 1.1. says very little).
- There are way too many typos, missing periods, line-initial periods and commas, sentences starting with lower-case characters, …

- As an example of the inconsistencies in the thesis, consider the bibliography:
    - inconsistent first names: initials for some, full names for others
    - inconsistent publishing date: *2009* vs *July 2004, 04 2009, sep 2020*
    - inconsistent page ranges: *page 2-4* vs *pages 2-4* vs *2-4*
    - inconsistent DOIs: url vs doi id vs both
    - wrong capitalization (charles university, ...)
    - the last work cited:
        - wrong year (2008 instead of 2018)
        - wrongly sorted
        - the author is *František Čermák*, but above, the same author is listed as *F. Čermák*
        - the article is in Czech: thus the name of the article should be in Czech, ideally with an English translation in parentheses; not in English only
        - the journal name misses diacritics
    - None of these issues are serious, but there are way too many of them. Moreover, it would take about ten minutes to fix them, and significantly improve the overall impression and readability.

Comments about the code

The code would benefit from:

- more comments, both docstrings and implementation comments,
- following PEP8 code style recommendations,
- unit tests, and
- type hints.

Prague, September 7, 2020
Jiří Hana