# Posudek diplomové práce

**Matematicko-fyzikální fakulta Univerzity Karlovy**

| | |
|---|---|
| **Autor práce** | Niyati Bafna |
| **Název práce** | Empirical Models for an Indic Language Continuum |
| **Rok odevzdání** | 2022 |
| **Studijní program** | Computer Science – Language Technologies and Computational Linguistics |

| | | | |
|---|---|---|---|
| **Autor posudku** | RNDr. Daniel Zeman, Ph.D. | **Role** | Oponent |
| **Pracoviště** | Ústav formální a aplikované lingvistiky | | |

**Text posudku:**

The thesis looks at cognate distribution in the dialect continuum of Indo-Aryan languages (plus one non-Indo-Aryan language) in northern India and surrounding areas. Many of the studied languages have little or no resources for language technology; therefore an important part of the work is finding and preparing data for these languages. Where permitted with respect to copyright, the author makes the processed corpora available for follow-up research; for the remaining cases, the author provides her processing pipeline so that others can obtain the data from the original source. This itself is an important contribution of the work to the research community.

The second part of the work is experimenting with various approaches to detection of cognates among these languages. While there is previous work on cognate detection, the author shows that her situation is quite different, as very little data is available, and there are no good seeds with known cognates. In addition, it is not easy to find data on which the cognate detection algorithm could be evaluated, and the available data have their own problems. This means that the results are not directly comparable to those known from literature on cognate detection.

There are 73 pages (plus Bibliography), of which 14 are the introduction, background and related work ; the remaining 59 describe author’s own work. The text is well organized and written in very good English with negligible number of typos. Throughout the text it is quite clear what has been done and why. I appreciate both the well explained reasoning about decisions taken (such as Section 5.3.1) and the discussion of the results of experiments. The discussion of related work and background literature seems more than sufficient to me.

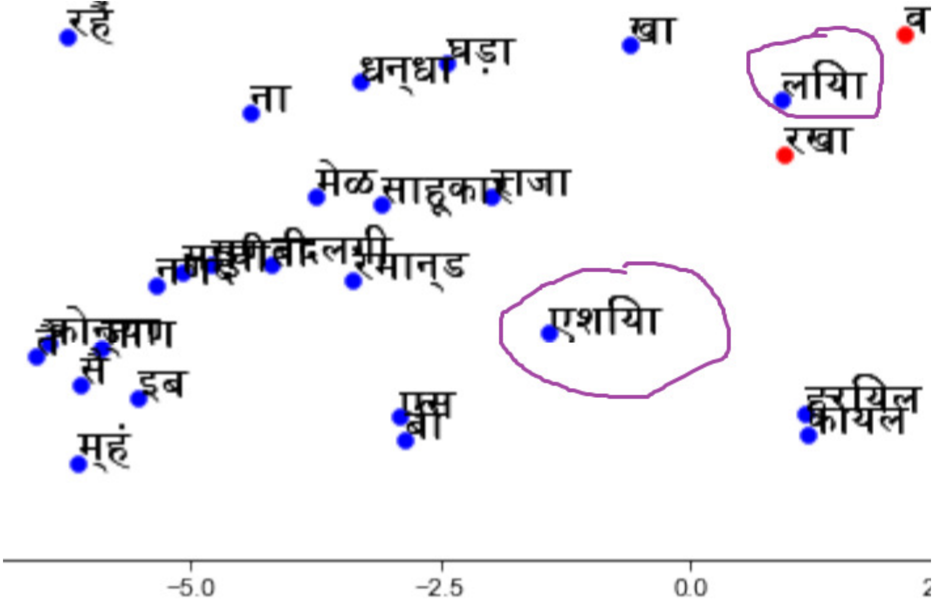
To summarize, I believe that the present thesis complies with the standards expected at the faculty and I recommend it to the defense.

## Specific Comments and Questions

- Page 9: “languages that we do not work with, such as Gujarati…” … This is a bit confusing, as the HinDialect data in Lindat contain Bengali, Gujarati, and Panjabi.
- Page 10, Table 1.1: I understand that the list of languages cannot be exhaustive but perhaps it would be interesting to include Kangri (Himachal Pradesh), which has a small Universal Dependencies treebank since May 2021.
- Pages 15 and 16: The running text seems to be interrupted on the page break between

these two pages.

- Page 23, Figure 4.2: My first question was, “Where is Sanskrit and Pali?” Later I found out that you explain it in the text, but since the figure appears early, it would be wise to also briefly explain it in the caption of the figure. More importantly, the provided reason for their exclusion is “since they are both dead languages”. But why would you want to exclude dead languages? Wouldn’t it be interesting to see how much lexical overlap the modern languages have with Sanskrit and Pali?
- Page 23, formula (4.2): What does “i” and “j” mean in this formula? Shouldn’t it be “1” and “2” instead?
- Page 28, formula (4.3): Where is “j”? In this form, $O_{ijc}$ would be 1 for any *ijc*.
- Page 28: For the subword experiments, are there any changes in filtering the low-frequency words? Could low-frequency words contribute to high-frequency subwords?
- As for Devanagari closely corresponding to pronunciation (footnote 2 on page 28 and before): Did you perform Unicode normalization of your data? For example, /z/ can be written as a single character “DEVANAGARI LETTER ZA”, or as a combination of “DEVANAGARI LETTER JA” and “DEVANAGARI SIGN NUKTA”, in both cases resulting in the glyph ज़. Is it guaranteed that the same glyph is always encoded the same way in the data?
- Page 37: “We treat corresponding/cognate words as parallel data…” … How do you do that when in fact your data is not parallel? How does the training data for the alignment algorithm look like?
- Page 38: “since we haven’t added horizontal edges between sisters” … Why haven’t you do that?
- Page 48, the measure *cl_integ*: Why should we be interested in the percentages of K nearest neighbors that belong to the opposite language?
- Page 49: It might be easier for the English-speaking reader if the word plots were Romanized. Furthermore, I find it hard to read some of them. There seem to be two vowel characters attached to the same consonant; could it be an issue with rendering Devanagari in the plot?



And if the upper-right word’s correct rendering is लिया *liyā*, why is it not colored as also belonging to Hindi-Urdu?

- Page 62, Table 8.3 (and same issue later with 8.4): What do the columns mean? Why is “Total in test” not the sum of “Unique in test” and “Common”?

- Page 63, regarding transliteration quality: Would the wrong transliteration always hamper cognate detection? Perhaps the wrongly transliterated word is still sufficiently similar to its Hindi counterpart?

**Práci doporučuji k obhajobě.**

**Práci nenavrhuji na zvláštní ocenění.**

*Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).*

**Datum** 25. srpna 2022 **Podpis**