Automated Cognate Discovery in the Context of Low-Resource Sami Languages

Eliel Soisalon-Soininen[1] and Mika Hämäläinen[2]

[1] Department of Computer Science
[2] Department of Digital Humanities
University of Helsinki

1 Introduction

The goal of our project is to automatically find candidates for etymologically related words, known as *cognates*, for different Sami languages. At first, we will focus on North Sami, South Sami and Skolt Sami nouns by comparing their inflectional forms with each other. The reason why we look at the inflections is that, in Uralic languages, it is common that there are changes in the word stem when the word is inflected in different cases. When finding cognates, the non-nominative stems might reveal more about a cognate relationship in some cases. For example, the South Sami word for arm, *gïete*, is closer to the partitive of the Finnish word *kättä* than to the nominative form *käsi* of the same word.

The fact that a great deal of previous work already exists related to etymologies of words in different Sami languages [2, 4, 8] provides us with an interesting test bed for developing our automatic methods. The results can easily be validated against databases such as Álgu [1] which incorporates results of different studies in Sami etymology in a machine-readable database.

With the help of a gold corpus, such as Álgu, we can perfect our method to function well in the case of the three aforementioned Sami languages. Later, we can expand the set of languages used to other Uralic languages such as Erzya and Moksha. This is achievable as we are basing our method on the data and tools developed in the Giellatekno infrastructure [11] for Uralic languages. Giellatekno has a harmonized set of tools and dictionaries for around 20 different Uralic languages allowing us to bootstrap more languages into our method.

2 Related Work

In historical linguistics, cognate sets have been traditionally identified using the comparative method, the manual identification of systematic sound correspondences across words in pairs of languages. Along with the rapid increase in digitally available language data, computational approaches to automate this process have become increasingly attractive.

Computationally, automatic cognate identification can be considered a problem of clustering similar strings together, according to pairwise similarity scores given by some distance metric. Another approach to the problem is pairwise

classification of word pairs as cognates or non-cognates. Examples of common distance metrics for string comparison include edit distance, longest common subsequence, and Dice coefficient.

The string edit distance is often used as a baseline for word comparison, measuring word similarity simply as the amount of character or phoneme insertions, deletions, and substitutions required to make one word equivalent to the other. However, in language change, certain sound correspondences are more likely than others. Several methods rely on such linguistic knowledge by converting sounds into sound classes according to phonetic similarity [**?**]. For example, [15] consider a pair of words to be cognates when they match in their first two consonant classes.

In addition to such heuristics, a common approach to automatic cognate identification is to use edit distance metrics using weightings based on previously identified regular sound correspondences. Such correspondences can also be learned automatically by aligning the characters of a set of initial cognate pairs [3, 7]. In addition to sound correspondences, [14] and [6] also utilise semantic information of word pairs, as cognates tend to have similar, though not necessarily equivalent, meaning. Another method heavily reliant on prior linguistic knowledge is the LexStat method [9], requiring a sound correspondence matrix, and semantic alignment.

However, in the context of low-resource languages, prior linguistic knowledge such as initial cognate sets, semantic information, or phonetic transcriptions are rarely available. Therefore, cognate identification methods applicable to low-resource languages calls for unsupervised approaches. For example, [10] address this issue by investigating edit distance metrics based on embedding characters into a vector space, where character similarity depends on the set of characters they co-occur with. In addition, [12] investigate several unsupervised approaches such as hidden Markov models and pointwise mutual information, while also combining these with heuristic methods for improved performance.

3 Corpus

The initial plan is to base our method on the nominal XML dictionaries for the three Sami languages available on the Giellatekno infrastructure. Apart from just translations, these dictionaries contain also additional lexical information to a varying degree. The additional information which might benefit our research goals are cognate relationships, semantic tags, morphological information, derivation and example sentences.

For each noun the noun dictionaries, we produce a list of all its inflections in different grammatical numbers and cases. This is done by using a Python library called Uralic NLP [5], specialized in NLP for Uralic languages. Uralic NLP uses FSTs (finite-state-transducers) from the Giellatekno infrastructure to produce the different morphological forms.

We are also considering a possibility of including larger text corpora in these languages as a part of our method for finding cognates. However, theses languages

have notoriously small corpora available, which might render them insufficient for our purposes.

4 Future Work

Our research is currently at its early stages. The immediate future task is to start implementing different methods based on the previous research to solve the problem. We will first start with edit distance approaches to see what kind of information those can reveal and move towards a more complex solution from there.

A longer-term future plan is to include more languages into the research. We are also interested in a collaboration with linguists who could take a more qualitative look at the cognates found by our method. This will nourish interdisciplinary collaboration and exchange of ideas between scholars of different backgrounds.

We are also committed to releasing the results produced by our method to a wider audience to use and profit from. This will be done by including the results as a part of the XML dictionaries in the Giellatekno infrastructure and also by releasing them in an open-access MediaWiki based dictionary for Uralic languages [13] developed in the University of Helsinki.

References

1. Álgu-tietokanta. saamelaiskielten etymologinen tietokanta (Nov 2006), http://kaino.kotus.fi/algu/
2. Aikio, A.: The Saami loanwords in Finnish and Karelian. Ph.D. thesis, University of Oulu, Faculty of Humanities (2009)
3. Ciobanu, A.M., Dinu, L.P.: Automatic detection of cognates using orthographic alignment. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 99–105 (2014)
4. Häkkinen, K.: Suomen kirjakielen saamelaiset lainat. Teoksessa Sámit, sánit, sátnehámit. Riepmočála Pekka Sammallahtii miessemánu 21, 161–182 (2007)
5. Hämäläinen, M.: UralicNLP (Jan 2018), https://doi.org/10.5281/zenodo.1143638, doi: 10.5281/zenodo.1143638
6. Hauer, B., Kondrak, G.: Clustering semantically equivalent words into cognate sets in multilingual lists. In: Proceedings of 5th international joint conference on natural language processing. pp. 865–873 (2011)
7. Kondrak, G.: Identification of cognates and recurrent sound correspondences in word lists. TAL 50(2), 201–235 (2009)
8. Koponen, E.: Lappische lehnwörter im finnischen und karelischen. Lapponica et Uralica. 100 Jahre finnisch-ugrischer Unterricht an der Universität Uppsala. Vorträge am Jubiläumssymposium 20.–23. April 1994 pp. 83–98 (1996)
9. List, J.M., Greenhill, S.J., Gray, R.D.: The potential of automatic word comparison for historical linguistics. PloS one 12(1), e0170046 (2017)
10. McCoy, R.T., Frank, R.: Phonologically informed edit distance algorithms for word alignment with low-resource languages. Proceedings of the Society for Computation in Linguistics (SCiL) 2018 pp. 102–112 (2018)

11. Moshagen, S.N., Pirinen, T.A., Trosterud, T.: Building an open-source development infrastructure for language technology projects. In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16. pp. 343–352. No. 85, Linkping University Electronic Press; Linkpings universitet (2013)
12. Rama, T., Wahle, J., Sofroniev, P., Jäger, G.: Fast and unsupervised methods for multilingual cognate clustering. arXiv preprint arXiv:1702.04938 (2017)
13. Rueter, J., Hämäläinen, M.: Synchronized mediawiki based analyzer dictionary development. In: Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages. pp. 1–7 (2017)
14. St Arnaud, A., Beck, D., Kondrak, G.: Identifying cognate sets across dictionaries of related languages. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2519–2528 (2017)
15. Turchin, P., Peiros, I., Murray, G.M.: Analyzing genetic connections between languages by matching consonant classes. Vestnik RGGU. Seriya "Filologiya. Voprosy yazykovogo rodstva", (5 (48)) (2010)