The Journal of Open Source Software

Omorfi: Open morphology of Finnish

Tommi A Pirinen[1]

1 University of Hamburg

DOI: 10.21105/joss.02184

Software

- Review
- Repository
- Archive

Editor: Hugo Ledoux

Reviewers:

- @desilinguist

Submitted: 25 February 2020
Published: 15 May 2020

Summary

Finnish is a natural language spoken mainly in Finland by ca. 6 million people as of 2020. Unlike many European languages, it has rich morphological system, which makes it slightly more complicated to use in computational systems, for example, a single word may be inflected in thousands of acceptable word-forms, making naive list- or statistics-based language models less effective. In omorfi, we solve this problem by generating inflecting dictionaries to be used in conjunction with finite-state automata-based morphological analysis software (Beesley, K.R. and Karttunen, L., 2003), available as open source from (Lindén et al., 2013).

Omorfi project consists of few distinct parts:

1. *a lexical database* containing hundreds of thousands of words, curated by langauge experts from open source repositories, such as wiktionary
2. a collection of *scripts* to convert lexical database into formats used by upstream NLP tools
3. an *autotools setup* to build, install, package, deploy or test the system easily
4. a collection of relatively bindings and libraries to programming languages popular within NLP community, such as python, Java and C++

The omorfi project is designed to be used by computer scientists who need a reliable model of Finnish language for computational natural language understanding software, or computational linguists wishing to study the grammatical features of Finnish. It has been used to implement various natural language processing software including spell-checking and correction, machine translation and dependency syntax treebanking.

Background

Omorfi has been developed since 2008, and is relatively stable lexical database for Finnish. It aggregates lexical data from various open source word-lists of Finnish into one big database:

- Nykysuomen sanalista (LGPL),
- Joukahainen (GPL) and
- FinnWordNet (Princeton Wordnet licence / GPL; relicenced with kind permission from University of Helsinki)
- Finnish Wiktionary and
- English Wiktionary (Creative Commons Attribution–ShareAlike).

Some words have also been collected by omorfi developers and contributors and are GPLv3 like the rest of the software package. All words have been manually verified by omorfi contributors.

Omorfi has been used in number of research projects and applications, some of which are listed in the official documentation under articles. Few of the more notable recent research include: machine translation (Rubino et al., 2015), OCR (Silfverberg & Rueter, 2015), semantic web (Mäkelä, 2014), named entity resolution (Ruokolainen, Kauppinen, Silfverberg, & Lindén, 2019) and spell-checking and correction (Pirinen, 2014).

Functionalities

Omorfi provides the functionalities directly through the underlying morphological analysis engines, such as (Lindén et al., 2013), but also through APIs for popular programming languages such as python, java and C++. For less advanced users, another project (Hämäläinen, 2019), provides user-friendly interfaces.

The interfaces provide all the main functionalities of natural language processing pipelines: given a string containing Finnish text, the text can be tokenised into words and punctuation or similar units, for each token a list of potential analyses can be retrieved, as well as potential morphological segmentations, hyphenations or spelling corrections. For example, given a string: `tässä on "sanoja."`, the tokenisation may return a list of tokens: `[tässä, on, ", sanoja, ., "]`, analysis of the first token might return a set of potential analyses: `{tämä PRON|PronType=Dem|Number=Sing|Case=Ine, tässä ADV}`, whereas the morphological segmentation would produce set of potential answers: `{tä ssä, tässä}`. The exact representation may vary between programming languages and API versions.

Omorfi contains a continuous integration test suite to guarantee the quality of lexical database after each contribution is integrated, this is especially crucial since contributions from wiktionary imported at regular intervalls.

Acknowledgements

We acknowledge all the professors, teachers and students who have been instrumental to creation and continued development of the system, for up-to-date list please view the THANKS file included with the system.

In chronological order of the project history:

- Inari Listenmaa
- Kimmo Koskenniemi
- Krister Lindén
- Erik Axelson
- Miikka Silfverberg
- Anssi Yli-Jyrä
- Inari Listenmaa
- Sjur Moshagen
- The students and staff ast Uni. Helsinki for extensive testing.
- University of Turku bio-NLP group (a lot of resources have been exchanged back and forth)
- Research institute of languages in Finland for Finnish wordlists
- Joukahainen / Voikko contributors
- GF contributors
- fi.wiktionary.org contributors
- Universal Dependencies (Finnish) contributors
- Unimorph contributors
- en.wiktionary.org contributors

References

Beesley, K.R. and Karttunen, L. (2003). *Finite State Morphology* (p. 503). CSLI publications.

Hämäläinen, M. (2019). UralicNLP: An nlp library for uralic languages. *Journal of Open Source Software*, *4*(37), 1345. doi:10.21105/joss.01345

Lindén, K., Axelson, E., Drobac, S., Hardwick, S., Kuokkala, J., Niemi, J., Pirinen, T. A., et al. (2013). HFST a system for creating NLP tools. In *International workshop on systems and frameworks for computational morphology* (pp. 53–71). Springer. doi:10.1007/978-3-642-40486-3_4

Mäkelä, E. (2014). Combining a rest lexical analysis web service with sparql for mashup semantic annotation from text. In V. Presutti, E. Blomqvist, R. Troncy, H. Sack, I. Papadakis, & A. Tordai (Eds.), *The semantic web: ESWC 2014 satellite events* (pp. 424–428). Cham: Springer International Publishing.

Pirinen, T. A. (2014). *Weighted finite-state methods for spell-checking and correction* (PhD thesis). University of Helsinki.

Rubino, R., Pirinen, T. A., Espla-Gomis, M., Ljubešić, N., Rojas, S. O., Papavassiliou, V., Prokopidis, P., et al. (2015). Abu-matran at wmt 2015 translation task: Morphological segmentation and web crawling. In *Proceedings of the tenth workshop on statistical machine translation* (pp. 184–191).

Ruokolainen, T., Kauppinen, P., Silfverberg, M., & Lindén, K. (2019). A finnish news corpus for named entity recognition. *Language Resources and Evaluation*, 1–26.

Silfverberg, M., & Rueter, J. (2015). Can morphological analyzers improve the quality of optical character recognition? In *Septentrio conference series* (pp. 45–56).