Analysing Finnish with word lists: The DDI approach to morphology revisited

Atro Voutilainen
FIN-CLARIN
University of Helsinki
atro.voutilainen@helsinki.fi

Maria Palolahti
FIN-CLARIN
University of Helsinki
maria.palolahti@helsinki.fi

Abstract

Morphological lexicons for morphologically complex languages provide good text coverage at the cost of overgeneration, difficulty of modification, and sometimes performance issues. Use of simple, manageable lexicon forms – especially lists – for morphologically complex languages may appear unviable because the number of possible word-forms in a morphologically complex language can be prohibitively high. We created and experimented with a list-based lexicon for a morphologically complex language (Finnish), and compared its coverage with that of a mature morphological analyser on new text in two experimental settings. The observed smallish difference in coverage suggests the viability of using simple and easy-to-modify list-based lexicons as an initial part of morphological analysis, to increase developer control on the vast majority of input tokens.

Tiivistelmä

Morfologiset leksikot morfologisesti kompleksisille kielille mahdollistavat korkean kattavuuden käytettäessä morfologista analysaattoria tekstien analyysiin. Toisaalta täysimittaiset morfologiset leksikot tuottavat toivottujen analyysien lisäksi paljon semanttisesti outoja analyyseja. Lisäksi morfologisen leksikon jatkokehittäminen haluttua sovellusta varten edellyttää parhaassakin tapauksessa huolellista ja työlästä perehtymistä morfologiseen kuvaukseen ja kehitysympäristöön. Listamuotoinen leksikko olisi yksinkertainen ja helppo muokata, ja siksi periaatteessa soveltajaystävällisempi vaihtoehto morfologiselle leksikolle. Listamuotoista leksikkoa voidaan pitää kuitenkin epätodennäköisenä vaihtoehtona morfologiselle leksikolle, koska esimerkiksi suomen morfologia (runsas taivutus, johto-oppi ja yhdyssananmuodostus) mahdollistavat suomen sananmuotojen erittäin korkean määrän. Tässä artikkelissa esittelemme kokeiluja, joissa olemme luoneet listapohjaisen leksikon suomen kielelle ja vertailleet sen kattavuutta kypsän morfologisen analysaattorin kattavuuteen kahdella koejärjestelyllä. Havaittu ero kattavuudessa on melko pieni, mikä tukee oletusta listapohjaisen leksikkomuodon käyttökelpoisuudesta morfologisesti kompleksisen kielen käsittelyssä.

Proceedings of the 4th International Workshop for Computational Linguistics for Uralic Languages (IWCLUL 2018), pages 171–181,
Helsinki, Finland, January 8–9, 2018. ©2018 Association for Computational Linguistics

1 Introduction

In NLP, a text analysis pipeline usually contains a basic component for lexical analysis: provision of a lexical analysis (or several, in case of lexical ambiguity) to tokens (word-like units). The knowledge base used by a lexical analyser can consist of a long (but simple) list of tuples (e.g. word-form, lemma, tags for POS and inflection) or of a complex morphological lexicon (lexical entries, inflectional or derivational morphemes, rules for combining these to account for correspondences between surface forms and lexical forms and rules for adding appropriate grammatical tags to the lexical analyses).

In a festschrift to a notable researcher in finite state morphology, Ken Church (2005) somewhat provocatively argues for a DDI (”don't-do-it”) approach to morphology: though traditionally a practical memory-sparing necessity for morphologically complex languages, lexical analysis with rule-based morphological lexicons tends to produce, as a side effect, spurious analyses that compromise the utility of the NLP pipeline in practical applications. Church gives examples from text-to-speech synthesis, information retrieval, part-of-speech tagging and spelling correction as support for his argument for a simple list-based lexical analysis. In the absence of a list-based lexicon, the application designer may skip the use of a linguistic lexical component altogether in favour of a more simplistic technique, as Kettunen (2013) has shown in the case of an IR system.

The authors of this paper have worked with linguistic models for Finnish NLP (morphology, tagging, syntax) in the symbolic/linguistic (rather than statistical/ML) paradigm. Though we are sceptical about adopting the DDI approach as such to morphology or other levels of linguistic analysis, we accept that there is a grain of truth in Church's argument about morphology: use of a full-fledged morphological lexicon for analysing a morphologically complex language can compromise developer control over the resulting analysis. Modifying a complex morphological lexicon for satisfactory analysis from the application point of view may be unrewarding even for an experienced linguist; for those inexperienced in morphology (i.e. most of the application builders) the only options may be either using the morphological lexicon as such (with all its undesirable side effects) or looking for other solutions to replace linguistic components entirely.

Methods to increase control over lexical analysis to facilitate successful integration in practical applications should be of interest to computational linguists, too. A list-based lexicon is arguably simple and easy to manipulate without the risk of unwelcome side effects. Should a list-based lexicon work on a morphologically complex language (in this case, Finnish) with a reasonable coverage, inclusion of a list-based, easy-to-manage lexicon (e.g. as a first part of morphological analysis) might be a user-friendly option to increase usability of an NLP pipeline in an application.

We are not aware of studies on attempting to generate and evaluate extensive list-based lexicons for morphologically complex languages. In this paper, we report generation of a large list-based lexicon for Finnish, and compare its performance to that of a mature linguistic morphological analyser in the analysis of new text. We also report a part-of-speech tagging experiment with the two alternative lexical analysers to get some data on how the use of a list-based lexicon affects tagging accuracy.

Next, we review some data on Finnish morphology and lexicons and consider options to generate a list-based lexicon.

2 Issues with a morphological lexicon

A full-scale morphological lexicon for a morphologically complex language has the desirable property for the application developer in that it enables recognition and analysis of a high percentage of the word-forms of the language, though the language has a very large number of potential word-forms. Unfortunately, there are also some features related to morphological lexicons and their development or maintenance that make morphological lexicons less desirable for application developers.

- Overgeneration. Even though a mature morphological analyser provides a correct and useful analysis to most of its input, full account for inflection, derivation and compounding in the morphological grammar also tends to result in semantically/ontologically spurious analyses, use of which is likely to compromise application performance. As an example, here is a morphological analysis for the Finnish sentence *Lisäaineisiin kuuluu niin askorbiinihappo kuin myös beetakaroteenikin.* (Additive substances include not only ascorbine acid but also beta-carotene):

```
"<Lisäaineisiin>"
        "lisäaine" N Pl Ill #2
        "lisäaineinen" A Pl Ill #2
        "lisäaineisi" N Sg Ill #3
        "lisäaineisä" N Pl Ill #3
"<kuuluu>"
        "kuulua" V Act Ind Pres Sg3 #1
        "kuuluu" Adv #1
        "kuuluu" N Sg Nom #2
"<niin>"
        "ne" Pron Dem Pl Ins #1
        "niin" Adv Dem #1
        "niin" Adv #1
        "niin" CCM #1
"<askorbiinihappo>"
        "askorbiinihappo" N Sg Nom #2
"<kuin>"
        "kui" N Pl Ins #1
        "kui" N Sg Gen #1
        "kuin" Adv #1
        "kuin" CC #1
        "kuin" CS #1
        "kuu" N Pl Ins #1
"<myös>"
        "myödä" V Act Imprt Sg2 S #1
        "myös" Adv #1
        "myös" CC #1
"<beetakaroteenikin>"
        "beetakaroteeni" N Sg Nom Kin #2
"<.>"
        "." Pun
```

Along with conventional analyses, the Omorfi analyser (Pirinen, 2015) also provides rather implausible alternatives that challenge downstream processing, such as:

  - side effects of compounding: ”lisäaineisi” (additive substance daddy), ”lisäaineisä” (additive substance father), ”kuuluu” (moon bone)
  - side effects of inflection: ”ne” (by means of those), ”kuu” (by means of moons).
  - inclusion of non-standard Finnish in the lexicon, e.g. spoken and archaic varieties: ”kui” (how), ”myödä” (sell)

- Complexity. Developers of morphological lexicons usually are fully aware of the problems of overgeneration, and make efforts to keep overgeneration in control without too heavily sacrificing the recognition rates. As morphological lexicons for languages like Finnish tend to be complex in any case, fixing encountered problems in the morphological lexicon is probably not an option for the casual application developer: uninformed changes to the organisation of the lexical classes are likely to produce undesired side effects in other parts of morphological analysis.

- Performance issues. Though morphological lexicons are typically run with machines that use finite-state technology known for its efficiency, the resulting morphological analysers are not necessarily particularly competitive in terms of analysis speed. The performance cuts may result from the excessive size of the finite-state automata as well as from use of external processing to circumvent morphology-internal management limitations.

- Lack of standard. If the morphological lexicon is developed or maintained (in an open-source environment) without a strict adherence to a well-documented standard, there is also the risk that an update to the morphological lexicon contains undocumented changes to some mid- or high-frequency lexical classes or morphology that silently change subsequent processing results for the worse.

Given that there is a management/control problem with complex morphological lexicons, there is a need for a simple, manageable solution, such as the lexicon as an enumeration of word-forms with their lemmas and morphology. With improvements in computing resources, the list-form lexicon – even for a morphologically complex language – may be an option, as Church (2005) actually suggests.

How should a list-form lexicon for a language like Finnish be used? Church argues for lists as a stand-alone component for lexical analysis (no morphology is needed). Our view is less extreme: also morphological lexicons are useful and needed, e.g. to support creation of an initial (unedited) list-based lexicon, and to provide an analysis to tokens not recognised by the list lexicon.

The main question so far is, whether it is an option in the first place to generate a useful list-based lexicon for a morphologically complex language like Finnish. Koskenniemi (2013) provides some well-known statistics about Finnish:

- The inflectional system in Finnish morphology is complex. Each Finnish noun has about 2,000 inflections; each adjective, 6,000; each verb, close to 20,000.

- A rich derivational morphology as well as a fairly liberal compounding mechanism takes the complexity to much higher levels.

- Given a lexicon with a moderate number of basic lexical entries (a few hundred thousands rather than millions) and an artificial limitation to four-part compounds,[1] the number of legitimate word-forms in Finnish is already a septillion (10e24).

Technically, a list-based lexicon this long could perhaps be generated using a morphological lexicon as a word-form generator, but this is not a practical option. Our contribution is to show

- that a raw (unedited) list-based lexicon for a morphologically complex language (Finnish) focusing on actually-occurring word-forms in text corpora can be made with a mature morphological analyser
- and that the resulting list-based lexicon can be used to provide a high text coverage, if not quite as high as that available from use of a full morphological analyser.

Next, we report compilation of a list-based lexicon for Finnish by using text corpora and the Omorfi morphological analyser. We can view this automatically generated list as a ”raw” list lexicon that could serve as a starting-point for modifications (addition of new information, deletion of unwanted analyses, etc.) needed for adapting lexical analysis to further uses. Then, we report comparison of the recognition rate of the resulting raw list-based lexicon with that of the Omorfi analyser itself on new text (including a comparison from a POS tagging perspective).

Finally, we discuss whether this kind of corpus-oriented list-based lexicon reaches an interesting recognition rate to serve as a basis for further work. Our aim in this paper is not to go into the kinds of modification potentially needed for adapting lexical analysis to an application or another; instead, the raw list lexicon is made publicly available with the publication of the IWCLUL proceedings in ACL Anthology.

3 Generation of list-based lexicon

3.1 Method

Freely available collections of Finnish text were downloaded from the Web; sentence extraction and tokenisation was performed; a word list was generated from the tokenised sentences (even tokens that occurred only once in the corpus were included). The word-list was analysed with the Omorfi morphological analyser; the analysed tokens were submitted to non-contextual disambiguation for pruning out analyses with more compound boundaries (“#1” for non-compounds, “#2” for two-part compounds, etc.) than an alternative analysis for the token in question has. The tokens with the compound-wise simplest analyses were converted into a list.

For example, the word-form *edustavien* is analysed by Omorfi as three-ways ambiguous (the first two are non-compounds - a participle and an adjective for ”representative”; the last one is a compound noun *edus* (frontside) *tavi* (common teal):

```
edustavien
        "edustaa" V Act PcpVa Pl Gen #1
        "edustava" A Pl Gen #1
        "edustavi" N Pl Gen #2
```

[1] five- and six-part compounds are not very uncommon either

In this case, a spurious reading can be safely discarded with this heuristic (assuming most lexicographers would reject an entry for the Finnish equivalent of frontside common teal). The first two readings are then converted into entries for inclusion in the list lexicon, e.g:

```
edustavien~"edustaa" V Act PcpVa Pl Gen
edustavien~"edustava" A Pl Gen
```

3.2 Corpus data

The downloaded corpora from which the tokens were extracted were the following:

- Finnish Wikipedia (fiwiki*pages-articles.xml.bz2)
- EUBookshop corpus for Finnish, from the Opus corpus (Tiedemann, 2012)
- Europarl corpus (Koehn, 2005)
- Suomi24 corpus (unmoderated Finnish-language discussion forum, containing a large amount of informal Finnish and typos)
- FiWaC corpus (Ljubešić et al. 2016)

In all, the extracted sentences contain close to 3 billion tokens.

3.3 List lexicon

The resulting raw list lexicon contains 9.74 million entries for all parts of speech (file size: 443MB). Compared with the number of entries in a morphological lexicon, a list lexicon of ten million entries is very large. Compared with the estimated number of potential word-forms in Finnish (a septillion, see above) ten million is almost nonexistent.

4 Evaluation 1: coverage of lexical analysers

4.1 Method

The test texts were tokenised by a tokeniser for Finnish before submitting them to the lexical analysers used in the comparison. This enables identical tokenisation and easier comparison of the lexical analysers without compromising performance of either analyser. The tokenised texts were then submitted to lexical analysis. Coverage rates (percentage of tokens analysed for each lexical analyser) were calculated. The tokens that received an analysis only from the morphological analyser (but not from the list-based analyser) were extracted, counted and classified into compounds and non-compounds (most of the tokens without analysis were compound nouns).

4.2 Analysers

As morphological analyser, we used the freely available Omorfi morphological lexicon (Pirinen, 2015) in connection with the HFST package (Lindén et al. 2009). Omorfi is a wide-coverage mature lexicon and morphological grammar that has been developed and refined for several years. The morphological description for Finnish closely follows the state-of-the-art descriptive grammar *Iso suomen kielioppi* (Hakulinen et al. 2004).

The list-based Finnish lexicon was run with a simple Perl program . It takes about 19 seconds for the Perl program to parse the large list-based lexicon; lexical analysis itself is reasonably fast, based on the data structure used.

4.3 Test data

The test data consist of news articles and columns from YLE (Finland's national public service broadcasting company) and OKM (Ministry of education and culture). In all, the test data contain 25,503 tokens. The data were shuffled at the sentence level for copyright reasons.

4.4 Results from evaluation 1

- Recognition rates. The Omorfi analyser gave an analysis to 98.8% of the tokens (25,193 tokens out of 25,503). The list-based analyser gave an analysis to 97.1% of the tokens (24,772 tokens out of 25,503).
- Differences. There were 421 tokens in the test data that received an analysis from Omorfi but not from the list-based analyser. Of these 421 tokens, 359 (85.3%) are compounds (compound nouns for the most part). As a point of comparison, only 4.5% (1150) of the tokens in the whole test corpus were compounds. The compounding mechanism seems to be the most important source of gaps in the coverage of the list-based lexicon, relative to the morphological lexicon.
- Speed of lexical analysis on a HP Elitebook laptop (Intel Core i5-4300U CPU @ 1.90GHz × 4, with 15.3 GiB of memory) with Ubuntu Linux. Omorfi: about three thousand tokens per second. List analyser: about 1.5 million tokens per second.

5 Evaluation 2: morphological disambiguation with lexical analysers

In this second evaluation, we looked at how the use of a list-based lexicon affects performance of a linguistics-based constraint tagger on the test text used in the previous evaluation.

5.1 Grammars

The grammars run on the morphologically analysed sentences were written by Maria Palolahti as a part of an ongoing project, the documentation and results of which will be published later. The grammars are based on the Constraint Grammar framework (Karlsson et al. 1995); the parsing software used is vislcg3 (Bick and Didriksen, 2015).

Before ambiguity resolution proper, a local heuristic CG was applied for adding morphological analyses to tokens not analysed by the lexical analyser. In the CG

available at `http://scripta.kotus.fi/visk/etusivu.php`

available at `http://visl.sdu.dk/~eckhard/analyzer.pl` formalism, a typical APPEND rule adds a lemma and a morphological analysis to a token based on the form of the token itself and/or its local syntactic context. For instance, a token with an apparent genitive ending that is followed by a postposition may be analysed as a noun in the genitive. Specific APPEND rules are followed by default APPEND rules to ensure that all tokens get an analysis before disambiguation starts.

Morphological disambiguation is based on constraints that operate on a combination of lexical and morphological information. Constraints are grouped as subgrammars ordered on the basis of the linguistic phenomenon to be resolved and on the basis of their reliability. A mature CG typically contains a few thousand constraint rules that resolve a large majority of the ambiguity in the input with a low error rate, to make further levels of analysis and use feasible. The grammars used in the present experiment contain several thousand constraints.

5.2 Method

The two CGs were run in sequence on the outputs of the two lexical analysers. The disambiguated text versions were compared to each other using the Linux "sdiff" program. The differences were examined one by one by the first author. Those cases where only one of the systems produced a correct analysis were marked to indicate, which pipeline produced the correct analysis. The symbol "O|" indicates the pipeline with the Omorfi morphological analyser produced the correct analysis; "L|" indicates that the correct analysis was produced by the pipeline with the list-based lexical analyser.

```
List-based analyser               Morphological analyser (Omorfi)
Kaipaan       V_Act_Ind_Pres_Sg1 Kaipaan       V_Act_Ind_Pres_Sg1
valoa          N_Sg_Par           valoa          N_Sg_Par
,              Pun                ,              Pun
kevyitä        A_Pl_Par           kevyitä        A_Pl_Par
vaatteita      N_Pl_Par           vaatteita      N_Pl_Par
,              Pun                ,              Pun
torikahveja    N_Sg_Nom        O| torikahveja    N_Pl_Par
ja             CC                 ja             CC
pehmeiden      A_Pl_Gen           pehmeiden      A_Pl_Gen
iltojen        N_Pl_Gen           iltojen        N_Pl_Gen
vaivattomuutta N_Sg_Par           vaivattomuutta N_Sg_Par
.              Pun                .              Pun
```

For instance, in the above example sentence *Kaipaan valoa, kevyitä vaatteita, torikahveja ja pehmeiden iltojen vaivattomuutta* (I miss light, light clothes, coffee in the market place and the ease of soft evenings) the compound *torikahveja* (market coffees) was analysed differently by the two pipelines. The analysis by the Omorfi pipeline (Noun Plural Partitive) was marked as correct with the "O|" tag. The differences were then counted and analysed.

5.3 Results from evaluation 2

In the 25,503 tokens in the test data, there were 254 tokens that received a correct analysis from one tagging pipeline but not from the other. As can be expected, the differences were unequally divided:

- Of the differences, 220 were such that the pipeline with the Omorfi morphological analyser has the correct reading and the other pipeline with the list-based lexicon has not (i.e. the local CG that adds new lemmas and analyses to out-of-vocabulary words made a misprediction).

- Of the differences, 34 were such that the pipeline with the list-based analyser has the correct reading and the other pipeline with the Omorfi analyser has not.

In terms of analysis correctness, the pipeline with the Omorfi analyser thus has 186 (220 minus 34) fewer misanalyses than does the pipeline with the list-based lexical analyser (difference between the two pipelines: 0.7%).

The majority of the misanalyses resulted from an incorrect analysis by the local heuristic grammar. To a much smaller extent, there were also at least two other types of error:

- Domino effect: a token analysed correctly by both lexical analysers was disambiguated incorrectly due to misanalysis of a word in the context by the heuristic APPEND grammar

- Rule order: the two lexical analysers sometimes provide the alternative analyses to a token in a different order, which can affect the application order of CG disambiguation rules and result in different analyses (especially when there is a mispredicting disambiguation rule in the grammar).

6 Discussion and future work

We have shown that a simple operable list-based lexicon with a text coverage nearly equal to that of a morphological lexicon can be generated with a mature morphological analyser by focusing on actual tokens found in large text corpora (instead of attempting to enumerate all possible word-forms in the language). Given that modification of a morphological lexicon can be prohibitively difficult for an application developer, access to a list-based lexical component may provide substantial additional control over lexical analysis (and downstream NLP) to the application developer. We also observed a substantial analysis speed improvement when using the list-based lexicon.

Heuristic grammar-based analysis of word-forms in a morphologically complex language is a difficult task, which suggests that a morphological lexicon should be used on forms not represented in the list-based lexicon. In any case, generation of a high-quality list-based lexicon without a solid morphological lexicon and analyser would probably require a prohibitive amount of manual work. Bypassing linguistic morphology altogether (the DDI approach) does not seem justified by our experiments.

We have not addressed the question, what kinds of modifications could be made to a raw list-based lexicon to enable successful integration of a NLP pipeline in an application. Release of the raw list-based lexicon itself hopefully facilitates future experimentation.

Acknowledgements

We are grateful to two anonymous IWCLUL referees for useful comments on an earlier version of this paper. We are also grateful to IWCLUL participants for a lively post-

talk discussion, and in particular to Kimmo Koskenniemi for his solemn promise to make still another morphological lexicon for Finnish.

References

Eckhard Bick and Tino Didriksen, 2015. CG-3 — Beyond Classical Constraint Grammar. *Proc. NODALIDA 2015.*

Kenneth Church, 2005. The DDI Approach to Morphology. In Antti Arppe et al. (Eds.), *Inquiries into Words, Constraints and Contexts. Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday.* Gummerus.

Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen and Irja Alho, 2004. *Iso suomen kielioppi* [Big Finnish Grammar]. http://scripta.-kotus.fi/visk/etusivu.php. URN:ISBN:978-952-5446-35-7.

Fred Karlsson, Atro Voutilainen, Juha Heikkilä, Arto Anttila (Eds.), 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text.* Walter de Gruyter.

Kimmo Kettunen, 2013. Managing word form variation of text retrieval in practice – why language technology is not the only cure for better IR performance. *Proc. Trends in Information Management 2013.*

Philipp Koehn, 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *Proc. MT Summit 2005.*

Kimmo Koskenniemi, 2013. *Johdatus kieliteknologiaan, sen merkitykseen ja sovelluksiin* [Introduction to Language Technology, its significance and applications]. URL: http://hdl.handle.net/10138/38503. ISBN 978-952-10-8677-9.

Krister Lindén, Miikka Silfverberg and Tommi Pirinen, 2009. HFST tools for morphology - an efficient open-source package for construction of morphological analyzers. *International Workshop on Systems and Frameworks for Computational Morphology.*

Ljubešić, Nikola; Pirinen, Tommi and Toral, Antonio, 2016, *Finnish web corpus fiWaC 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/-11356/1074.

Tommi Pirinen, 2015. Omorfi - Free and open source morphological lexical database for Finnish. *Proc. NODALIDA 2015.*

Jörg Tiedemann, 2012. Parallel Data, Tools and Interfaces in OPUS. *Proc. LREC 2012.*