LEXICON-BASED MORPHOLOGICAL ANALYSIS OF LATVIAN LANGUAGE

Pēteris Paikens
University of Latvia, Institute of Mathematics and Computer Science (Riga, Latvia)

Abstract

This paper describes a practical solution for lexicon-based morphological analysis of Latvian language. As it is a flexive language, the core of this system is an implementation of word inflection based on a stem and its properties as listed in the lexicon. The main advantage of the described solution over similar implementations is augmenting the lexicon with methods for word derivation from related word stems, significantly increasing the recognition rate. The implemented system is able to provide full morphological detail for 96 % words of unrestricted Latvian language texts, even when using a rather limited lexicon of 25,000 word stems. For remaining unknown words, the system is extended with heuristics for recognising proper names, and determining verb and noun flexive forms based on ending, allowing a good quality guess for the linguistic properties of words that are not included in the lexicon. Such wide coverage allows the solution to be used in other linguistic tools as a transparent and robust layer for analysing word properties.

Keywords: morphology, part of speech, tagging, dictionary.

1. Introduction

For flexive languages, like Latvian language, morphological analysis and/or stemming often is the first required step in any text analysis process. However, there is a lack of publicly available morphological analysis tools for Latvian language, and most linguistic solutions and research – for example, current semantic projects in University of Latvia (Bārzdiņš et al 2007a) tend to use custom dictionaries to match exact word forms (instead of word stems) with the required information. Such approaches work acceptably within the domain covered by the dictionary, but for analysis of unrestricted corpora there will always be significant portion of words that are not included in the dictionary.

This paper describes a currently developed system that aims to provide a robust and extensible solution for morphological analysis for unrestricted corpora, in order to have an available solution to facilitate further analysis of Latvian language.

2. Solutions used currently

Most research purposes seem to use hard-coded dictionaries for the first steps of analysis, disregarding morphology entirely. Often the main reason for this is adaptation of tools originally designed for analysis of English language corpora, which don't

support flexive word forms, and treat all variations of a single stem as entirely separate words. This hampers linguistic analysis, but is often accepted due to technological difficulties. Availability of morphological tools for Latvian language might relieve this issue, and facilitate easy testing of other computational linguistics tasks on various corpora.

An interesting approach was seen (Krūze-Krauze 1998) that attempts to generate all possible words by defining formal grammar rules that govern the ways how different morphemes may combine together to make a single word. However, the system currently is in a dead-end, reaching an unmaintainable size of rules and special cases, and still fails to recognize a large portion of words, especially words adopted from other languages such as Greek or English.

There have also been a number of previous attempts of morphological analysis as well, with similar methods as described in the following section, but without much success, as the achieved coverage tends to be good enough for handmade research corpora, but is not acceptable for unrestricted text analysis. The current development attempts to include the experience of these previous attempts, especially in various heuristics for handling exceptional cases.

3. Lexicon-based analysis

In Latvian language, most word classes – nouns, verbs and adjectives – are flexive, consisting of an (almost) unchanging stem, and an ending that specifies various grammatical properties of the word. The exact endings vary depending on the basic stem, but almost all words in Latvian language can be split in a limited number (23 for the current implementation) of groups where every word form can be generated automatically. There are some irregular words, though (for example, *"būt" – "to be"*), but their number is rather limited, so they can be included manually in the lexicon.

Morphological analysis can be done by using this database of the endings in order to generate all possible variations where the ending is equal to the last letters of the analysable word, and looking up the remaining letters within the available dictionary of stems for a possible match. See Figure 1 for the workflow process. A major issue here is the stem changes that sometimes happen (Ceplīte et al 1991) in various word forms. This is analysed with a custom heuristics that lists all the cases occurring in literary Latvian language.

Ambiguity in this part of analysis is unavoidable, as there are many words in Latvian language where the part of speech details can be determined only in syntactic context. For example, word *"roku"* (given as an example in Figure 1) can be a word form of *"roks"* ("rock music"; masculine noun), various word forms in different cases of *"roka"* ("hand"; feminine nound), or a form of *"rakt"* ("to dig"; verb).

Thus the core data of the system is a dictionary of lexical units, containing word stems grouped in morphological types, and any information about these word stems that should be passed on in the results. As the goal of the system is to provide foundation for further syntactic and semantic analysis, a major focus is to provide extensibility for lexicon data, allowing the users to amplify the lexicon with any additional data relevant for the problem at hand. Current uses for additional information include verb transitivity information for the main lexicon, and semantic ontology groupings for a small research lexicon.
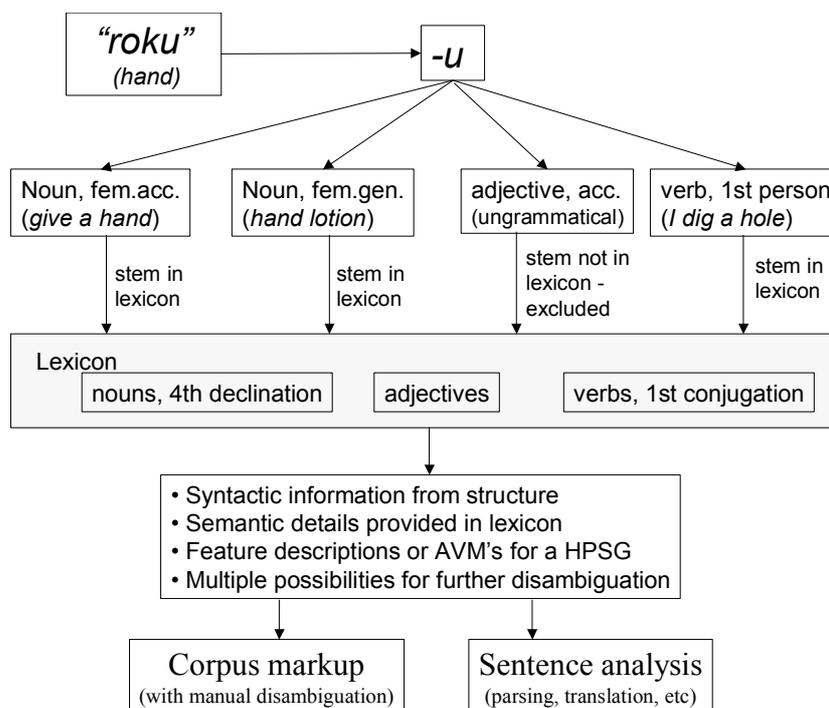
Figure 1 – Workflow of analysis process

4. Extending the lexicon

Of course, such analysis methods are limited by the extent of the dictionary used. This problem is further complicated by the fact that digitally available dictionaries in Latvian language are comparatively small. This research used a lexicon of approximately 27,000 word stems, based on an electronic version of an inverse dictionary (Soida, 1970). Initial analysis has showed that such a lexicon can cover 85–90% of unrestricted text[1]. This ratio is quite low and clearly not enough for practical purposes even with larger dictionaries. This has always been an issue for previous such developments, and was usually tackled by targeting a specific domain of text, and manually adapting the dictionary for that target domain.

However, analysis of missed words in these corpora (see Table 1) shows that most of them are in some way derived from other words that are included in our lexicon. This can be expected, as new words in Latvian language, as many others, often are formed by extending already existing words with similar meaning, or metaphorically using these words in a different meaning.

---

[1] Novels „Plāns ledus" and „Sofijas pasaule" („Sophie's world") were used as test corpora for this and other statistics.

Table 1 – Words not found in lexicon

Derivation via prefix	63%
Derivation via infix	2%
Compound nouns	2%
Proper nouns	5%
Reflexive verbs	11%
Not related to words in lexicon	15%
Erroneous words	2%

These derivation types can be all automatically analysed by attempting to match the word to stems in the existing lexicon. There is a rather small number of prefixes and infixes used in Latvian language, and it is computationally easy to try them all. Not all of these usages are valid, so this cannot be used for spell-checking type of solutions, but in the case when a new word is encountered in a corpus, and it matches some other in the lexicon with a derivation rule, we can be reasonably sure that we have encountered a correct (but new) word, and the appropriate part of speech information can be extracted.

On the other hand, if additional information is expected from the lexicon – for example, the inclusion of word senses in some ontology – then it is not safe to include this information. The derived words usually are related in meaning, but the nature of this relation may vary greatly, in some cases the relation is only metaphorical. Newly found words can then be used to enhance the lexicon, but this would require at least some review and approval for all non-morphological information.

5. Treatment of unrecognised words

The final part of this application, which gets used if the previous methods have failed to relate the word to some entry in our lexicon, is a heuristic for part-of-speech tagging based on the last letters of the analysable word.

A core part of the initial morphological analysis system is an exhaustive list of all endings for flexive word classes, linked with the morphological and part-of-speech information that each ending can represent. This list can also be used without looking at the lexicon – this will yield a lot of ambiguous possibilities, but still this process can usually exclude the majority of variations. For example, gender of the noun can usually be determined in this way, but several possibilities for number and case may remain. Many word forms in Latvian language can be uniquely identified in this way (Nau 1998), for example, verb participle forms. Comparison of the results with annotated corpus (Levāne 2001) shows that the generated variants always include the proper tagging – several mismatches were found in the comparison, but they were all found to be errors in the manual tagging.

For cases with more ambiguity, frequency distribution tables are used to determine the most likely word forms, or, alternatively, all possibilities are included, and ambiguity resolution can be left to the syntax part of analysis, since it is very likely that the specific form can be determined from grammatical requirements of case/number/gender agreement.

6. Technical implementation

The main consideration in the technical implementation has been compatibility with various other tools that may be used together with this morphological module. There are

two independent workflows implemented. One way of interaction is a batch-annotation mode, where a corpus is transformed into an XML file with the (ambiguous) morphological analysis results appended to each word. The other way is an online analysis service, which can be repeatedly queried for analysis of particular words, and can be accessed by other applications.

Current implementation is done in Java, with an interface that can be called from Prolog applications. Analysis performance is about 16,000 words per second on a standard desktop PC.

The lexicon and all other data are stored in XML files. For corpus tagging currently a custom XML format is used, but work is underway to use Tiger-XML formats everywhere to improve interoperability with different tools used in computational linguistics for other languages.

8. Evaluation and further developments

Current status of this development is stable enough to use it as a transparent layer of unrestricted text corpus analysis, extracting morphological and part-of-speech information for nearly all words within it. Increasing lexicon size will help decrease ambiguity of this analysis, but is not absolutely necessary. Using a modestly sized lexicon of 27,000 word stems and the abovementioned word derivation rules, 96 % of words in test corpora can be fully analysed, and for the remaining 4 % words all the part-of-speech possibilities are provided – which usually include two or three possibilities for grammatical case of the word.

Current usage of the system includes University of Latvia projects in semantic ontology (Bārzdiņš et al 2007b), and in corpora extraction from the Internet (Džeriņš et al 2007). The solution is considered to be publicly available for any research purposes upon contacting the author.

Details of all these tasks, naturally, can be developed further for improved results. In particular, further developments would include the following tasks:

- Adapting existing solutions for entity name recognition in order to treat proper nouns in a more precise way. This issue is common for many other languages, so solutions designed for English could be used.
- Changing the XML formats used to match Tiger-XML is underway.
- Support for morphological analysis of transliterated text[2], to improve coverage for corpora extracted from Internet. There are solutions that attempt to transform words from transliterated to proper form, but they are naturally ambiguous, and this ambiguity resolution can be improved if the morphological analysis is done directly on transliterated text.
- There are some ways of word derivation that occur less frequently and are not yet included – diminutive forms would be a good candidate for inclusion, as such forms occur in the language, but tend to be excluded from dictionaries.

---

[2] Spelling of text with exclusively latin characters, i.e. spelling *zaķīši* as *zakjiishi*, which is occasionally encountered on websites, blogs and Internet discussion forums.

12. References

Bārzdiņš, Guntis; Grūzītis, Normunds; Nešpore, Gunta; Saulīte, Baiba 2007a. Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order. In: Nivre, J.; Kaalep, H.; Muischnek, K.; Koit, M. (eds.) *NODALIDA 2007 conference proceedings*. Tartu: University of Tartu.

Bārzdiņš, Guntis; Grūzītis, Normunds; Levāne-Petrova, Kristīne; Nešpore, Gunta; Saulīte, Baiba 2007b. A deep discourse representation structure for theme-rheme and anaphora resolution. In: *Human Language Technologies 2007 conference proceedings*.

Ceplīte, Brigita; Ceplītis, Laimdots 1991. Latviešu valodas praktiskā gramatika. Rīga: Zvaigzne.

Džeriņš, Jānis; Džonsons, Kristaps 2007. Harvesting national language text corpora from the Web. In: *Human Language Technologies 2007 conference proceedings*.

Krūze-Krauze, Baiba 1998. Datorizēta latviešu valodas morfēmiski morfoloģiskā analīze. Rīga: Latvijas Universitāte.

Levāne, Kristīne 2001. Paula Bankovska romāna "Plāns ledus" pirmās nodaļas morfoloģiskā anotēšana un statistiskā analīze. Rīga: Latvijas Universitāte.

Nau, Nicole 1998. Latvian. Newcastle: LINCOM Europa.

Soida, Emīlija; Kļaviņa, Sarma 1970. Latviešu valodas inversā vārdnīca. Rīga: LVU.

PĒTERIS PAIKENS is a junior researcher of Institute of Mathematics and Computer Science, University of Latvia. He received his bachelor degree in computer science at University of Latvia. His research interests include corpus analysis, syntax analysis for Latvian language and formal grammar development.
E-mail: PeterisP@gmail.com.