Morphological analysis with limited resources: Latvian example

Pēteris PAIKENS, Laura RITUMA, Lauma PRETKALNIŅA
University of Latvia, Institute of Mathematics and Computer Science
peteris@ailab.lv, laura@ailab.lv, lauma@ailab.lv

ABSTRACT

We describe an approach for morphological analysis combining a rule-based word level morphological analyzer with statistical tagging, detailing its application to Latvian language. Latvian is a highly inflective Indo-European language with a rich morphology.

The tools described here include an implementation of Latvian inflectional paradigms, a morphological analysis tool with a guessing module for out-of-vocabulary words, and a statistical POS/morphology tagger for disambiguation of multiple analysis possibilities. Currently achieved accuracy with a training set of only ~40 000 words is 97.9% for part of speech tagging and 93.6% for the full morphological feature tag set, which is better than any previously publicly available taggers for Latvian.

We also describe the construction and methodology of the necessary linguistic resources – a morphological dictionary and an annotated morphological corpus, and evaluate the effect of resource size on analysis accuracy, showing what results can be achieved with limited linguistic resources.

KEYWORDS : morphology, inflective language, POS tagging, Latvian language, morphological corpus.

1 Introduction

For inflective languages, where a large part of grammatical meaning is expressed by the morphological features of words, a wide variety of computational tasks require a way to perform automated part of speech tagging and morphological analysis. It is needed both in specialized use cases such as linguistic research, and also in end-user tasks such as searching within documents or automated spelling correction.

Smaller languages usually have a limited amount of resources and effort available for developing linguistic resources such as annotated corpora or dictionaries, so it is useful to explore analysis methods that would work with smaller amount of resources and allow reusing software tools developed for other, larger languages.

In this paper we describe the construction process of such a toolkit for Latvian language, integrating word-level morphological analysis based on a formalization of Latvian inflection paradigms with a statistical tagger to exploit sentence context. We also provide an evaluation for the effect of resource (annotated corpora and lexicon) size on analysis accuracy.

2 Latvian Morphology

Latvian is an Indo-European language with around 1.5 million native speakers. It is a synthetic inflected language with rich morphology somewhat similar to the commonly analyzed Czech morphology (Hajič, 2000).

Latvian nouns and pronouns have 6 cases in singular and plural in traditional grammar. Nouns are traditionally divided in 6 declensions with different inflectional paradigms. Adjectives, numerals and some participles have 6 cases in singular and plural, 2 genders (masculine and feminine) and separate definite and indefinite forms. In verb conjugation system are 2 numbers (singular and plural), 3 persons, 3 tenses (present, future and past, both simple and compound) and 5 moods, as well as multiple types of participles. Qualitative adjectives and adverbs formed from these adjectives have also degrees of comparison noted in their word form.

The morphology creates more than 200 verb and participle forms derived from each verb lexeme, and more than 100 forms for each adjective. Many of the endings are overlapping, creating homoforms – for example, singular accusative and plural genitive forms are identical for many words.

2.1 Related Work on Latvian Morphological Analysis

The earliest experiments with automated Latvian morphological analysis have been performed in 1970s (Drīzule, 1978), implementing noun and adjective analysis. In 1990s, with the advent of personal computers, there have been multiple attempts to create analysis systems for all parts of speech (Greitāne, 1994; Levāne & Spektors, 2000; Sarkans, 1996, Vasiļjevs, Ķikāne & Skadiņš, 2004) based on linguistic rules for word endings and morphemes.

Systems currently being used in practice for Latvian morphology include lexicon based analysis systems (Paikens, 2007; Skadiņa, 2004) – while requiring more computational and dictionary resources, such systems provide better accuracy than earlier research.

Morphological analysis of Latvian is rather ambiguous – about half of words have multiple valid interpretations if viewed without context, so disambiguation as analyzed in this paper is an important open problem. There exists a recently developed morphological tagger based on Maximum Entropy Model (Pinnis and Goba, 2011), but it is not available to public[1].

3 Development of a Morphologically Annotated Corpus

A morphologically annotated corpus is a key resource for all further work – even for methods that do not require input from a large corpus, it is crucial to have at least a small set of verified data that can be used for testing and evaluation.

3.1 Morphological Annotation Standard

The morphological feature annotation standard used for Latvian corpora was initially (Levāne, 2000) derived from the annotation principles used for other languages in the MULTEXT-East project (Erjavec, 2004). It is a way to represent word annotation with a short tag, each character position representing a separate, independent feature. The meaning of each character position depends on the part of speech (marked in the first character) in order to keep the tag length short enough for human reading.

For an example, Figure 1 illustrates the morphological feature tag for noun *draugam*, the singular dative form of *draugs* (a friend).

Tagset for noun	part of speech	type	gender	number	case	declension
	n	**c**	**m**	**s**	**d**	**1**
	noun	common	masculine	singular	dative	first

FIGURE 1 – Example of a morphological tag for noun *draugam* ('friend') **ncmsd1**

It should be noted that in addition to purely morphological features, the annotation includes also lexical properties (such as type and declension in Figure 1) necessary for other research uses of the annotated corpora. The tag element names and values are matched to the ISOcat standard as recommended by CLARIN project[2].

The annotation process starts with generating the possible readings with an automatic analyzer described in the next section, and then a manual review and entry of missing

[1] The tagger is used in company Tilde proprietary tools - the training data and tagger are not available for other research purposes.

[2] http://www.clarin.eu

features. The speed of annotation is around 300 words per hour for a skilled operator with appropriate software tools.

3.2 Annotated Corpora

There have been multiple efforts on building morphologically annotated corpora of Latvian. Currently publicly available corpora are shown in Table 1. As noted earlier, the corpora were developed for projects of varying goals, and there are some differences between exact annotation standards used.

Corpus	Text source / domain	Tokens	Sentences
Balanced	Latvian Balanced Corpus[3]	50 795	3 940
Legal	EU documents[4]	23 359	1 038
Plāns Ledus	A fiction book	16 708	1 314
Latvijas Vēstnesis	A newspaper	28 956	2 035

TABLE 1 – Morphologically annotated Latvian corpora.

These corpora have been reviewed by a single annotator only. To ensure adequate data quality we performed a second annotator review and correction of the balanced corpus annotation to reduce the number of annotation errors, and serve as a valid 'gold standard' data for analyzer training and evaluation in this paper.

4 Automated Morphological Analysis

Our basic morphological analysis – generation of all possible morphological interpretations of a word form – is based on an earlier publicly available lexicon-based morphological analyzer (Paikens, 2007), extending it with additional lexical data. It is based on matching possible word form endings and the inflectional changes to stems as described in classical linguistic research, and verifies the stem candidates against a lexicon marked with declensions and conjugations of common nouns and verbs.

The currently used morphological lexicon has been assembled from multiple sources, including an electronic version of an inverse dictionary (Soida, 1970), manual review of the closed word classes and words with irregular inflection, scientific terminology data, and updates based on . It is not properly balanced – the contents reflect what resources were available, so coverage may vary depending on the text domain. The lexicon contains 47 000 lexemes.

Even with such lexicon size, 5-6% of test data is still out of vocabulary. Most of these words are formed according to Latvian grammatical rules, so it is still reasonable to deduce morphological properties based on the word ending, and for these cases, a 'guessing' system is implemented that generates a large number of possible analysis

[3] http://www.korpuss.lv

[4] White Paper. Preparation of the Associated Countries of Central and Eastern Europe for Inte-gration into the Internal Market of the Union.

options. This includes the correct reading for all except some 0.5-1% foreign words or brand names that are used literally as inflexive nouns, but happen to have an ending that matches a Latvian flexive form.

5 Statistical Disambiguation Methods

For many languages, pure morphological analysis will have a significant amount of ambiguity. For Latvian, our current analyzer gives multiple interpretations for 50-55% of words, with an average of 3.5-3.8 options for ambiguous words, depending on text domain, and similar amount of ambiguity has been observed in other morphologically rich languages (Yuret & Türe, 2006). The above ambiguity measurement includes morphological features – part of speech, case, number, gender, etc., and also lemmas in case of inflectional homonymity.

We examine two main use cases for disambiguation – choosing the most likely option for a single token, or selecting the most likely morphological tags for a whole sentence, looking at words in context. Single token analysis has less data for accurate disambiguation, but can be used in analysis of incomplete text fragments such as search queries, and is simpler to implement.

5.1 Baseline - Single Token Disambiguation

If there are multiple valid interpretations, clearly some of them are more frequent than others – we can intuitively note that some inflective forms may be more commonly used; or that one of theoretically possible lemmas is a rare, archaic word.

For this scenario, we can count the frequencies in a morphologically disambiguated corpus for two main features – the inflectional paradigm that generated the option, and the lexicon entry (if any) of the source lemma. This allows a quick estimation of the likelihoods, choosing the analysis option with the most likely paradigm and lexeme. While this method is naturally limited, it provides reasonable results with very tiny resources, providing us with a baseline to evaluate more complex options described later.

This is similar to the first stage of a Brill tagger if the surface form was seen in training corpus, but this heuristic generalizes well also to cases where the exact form was not seen before.

5.2 Morphological Tagging Within a Sentence

There are two main directions to use sentence context in disambiguation of homoforms in order to appy the appropriate morphological tags. One approach would be to invoke syntax rules, such as general syntactic analyzers (e.g. Bārzdiņš, Grūzītis, Nešpore & Saulīte, 2007 or Deksne & Skadiņš, 2011) that could also be adapted for morphological disambiguation. On the other hand, it is also possible to obtain these rules directly from an annotated corpus with machine learning algorithms. Our initial experiments with available Latvian syntactic analysers gave poor results due to limited syntactic coverage, driving us to the machine learning direction – although other research (Hajic,

Krbec, Kveton, Oliva & Petkevic, 2001; Hulden & Francom, 2012) suggests that a hybrid approach may bring further improvements.

Further in description we use our currently best performing solution, a conditional Markov model (CMM) based morphological tagging module. We have also trained various other systems, including hidden Markov model (HMM) and conditional random field (CRF) based classifiers, but we achieved better results with CMM.

The CMM module software is a modified version of the Stanford NLP[5] system CMM classifier implementation (Toutanova, Klein, Manning & Singer, 2003). A major difference between our solution and the original Stanford POS-tagger is the integration of the classifier with a rule based morphological analyzer supplying multiple possible analysis options to the classifier for disambiguation.

The standard approach for other languages (Hulden & Francom, 2012; Toutanova, Klein, Manning & Singer, 2003; Gahbiche-Braham, Bonneau-Maynard, Lavergne & Yvon, 2012) is to train a classifier on features directly derived from the word form string, such as letter n-grams, capitalization features, etc. While this may be effective for languages with a smaller range of word forms, this is not optimal for morphologically rich languages, as suggested by research in other languages (Youret & Türe, 2006). Word form specific features would greatly suffer from feature sparsity, as even in a huge training corpus many rarer word forms would not be seen at all; and a large part of word ending inflection rules cannot be adequately captured by letter n-gram features.

However, this morphological knowledge can be exploited by adding as training features the results from rule based morphological analysis described in section 4. That gives a reasonably accurate (contains correct form in 98% cases) list of what tags seem possible for each word. So in addition to the used classifier training features commonly used for other languages, we also supply a list of possible part-of-speech and tag options for the selected word and its closest neighbours. We also provide a 'recommended' POS and tag, calculated as described in section 5.1, which gives ~1% additional boost in accuracy. This change augments the machine learning of ending (letter n-gram) relations with morphological features with the linguistic rules in analyser, and allows to achieve good results with rather small training corpora.

6 Evaluation

6.1 Methodology

We used a morphologically annotated balanced corpus of 50 795 words, using 46 306 of it as training data (5 344 of it for tuning and developing the systems), and a separate set of 4 489 words for evaluation in this paper. Text content is taken as-is from the corpus, leaving intact any spelling issues or insertions of foreign words.

Lexical features such as declension, verb modality, semantic grouping, etc. are discarded for both training and evaluation data, as they can be retrieved afterwards from

[5] http://nlp.stanford.edu/software/tagger.shtml

the lexicon when the lemma is determined. The following morphological features are used for evaluation: part of speech, gender, number, case, person, verb mood, and definiteness for adjectives and participles.

6.2 Rule-based analyzer module evaluation

On our test corpus, the rule-based morphological analysis module includes a correct analysis option for 98.2% words, incorrect analysis for 1.3% words, and no analysis for 0.5% words (mostly insertions from other languages). Rule-based analysis results are unambiguous for 46.6% words, and the ambiguous words have on average 3.8 options each.

6.3 Statistical disambiguation methods

Comparing the results of automatic morphological disambiguation on the evaluation data set shows a tag accuracy level of 87.0% for baseline single token analysis and 93.6% for the best performing CMM model.

Both methods are suitable for analysis of large text corpora, with single token analysis being able to analyze approx. 100 000 words per second per core on a 2.8Ghz processor, and the CMM tagger around 3 000 words per second.

Reviewing the distribution of disambiguation errors by feature category (break-down shown in Table 2) indicates that the most common error is a combination of number and case mismatch, confusing singular accusative and plural genitive forms of nouns or whole noun phrases. These are homoforms for a large portion of Latvian nouns and adjectives, and both accusative and dative may be syntactically reasonable after a verb, indicating respectively the object or recipient of the action. We plan to reduce this class of errors by integration of morphological disambiguation with deeper syntactic analysis (statistical dependency parsers) that should be able to better resolve such ambiguities.

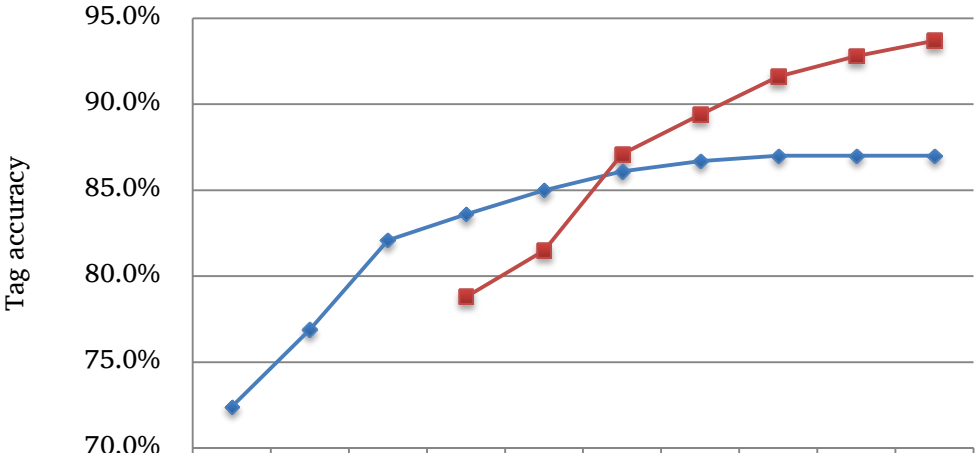
Part of speech	2.1 %
Gender	3.2 %
Number	4.5 %
Case	7.0 %
Verb mood	1.8 %
Person	0.8 %
Definiteness	1.4 %

TABLE 2 – CMM tagger error rates within feature categories.

6.4 Training data size effect on accuracy

Experiments on running the same disambiguation methods with limited training data, illustrated in Figure 2, show that the naive single token disambiguation quickly reaches its limit at around 10 000 words already. The CMM based model would likely provide

better accuracy with additional training data, which is also supported by experiments of Pinnis and Goba (2011) performed on a training set of 117 000 words, but it already provides an improvement above the single-token baseline with even very small training corpus such as 5 000 words.



	0	100	500	1 000	2 000	5 000	10 k	20 k	30 k	46 k
Single-token	72.4%	76.9%	82.1%	83.6%	85.0%	86.1%	86.7%	87.0%	87.0%	87.0%
CMM				78.8%	81.5%	87.1%	89.4%	91.6%	92.8%	93.7%

FIGURE 2 – Effect of corpus size on CMM disambiguation accuracy compared to single-token baseline

6.5 Effect of Lexicon Size on Accuracy

To evaluate the necessity of a morphological lexicon (a dictionary annotated with declensions or inflectional paradigms), we performed a series of tests, training and running the CMM classifier with an artificially reduced lexicon. The minimal dictionary contains 5 000 lexemes for the closed word classes – pronouns, conjunctions, prepositions, and irregular verbs, with further experiments measured by randomly adding nouns and verbs from the full dictionary up to the indicated limit.

The evaluation results shown in Figure 3 indicate that a proper lexicon has a strong impact in reducing error rate, however, when considering languages or dialects where large dictionaries are unavailable (such as the Latgalian language closely related to Latvian), it is not strictly necessary since our experiments show a practically usable accuracy of 90.6% even with the minimal lexicon.

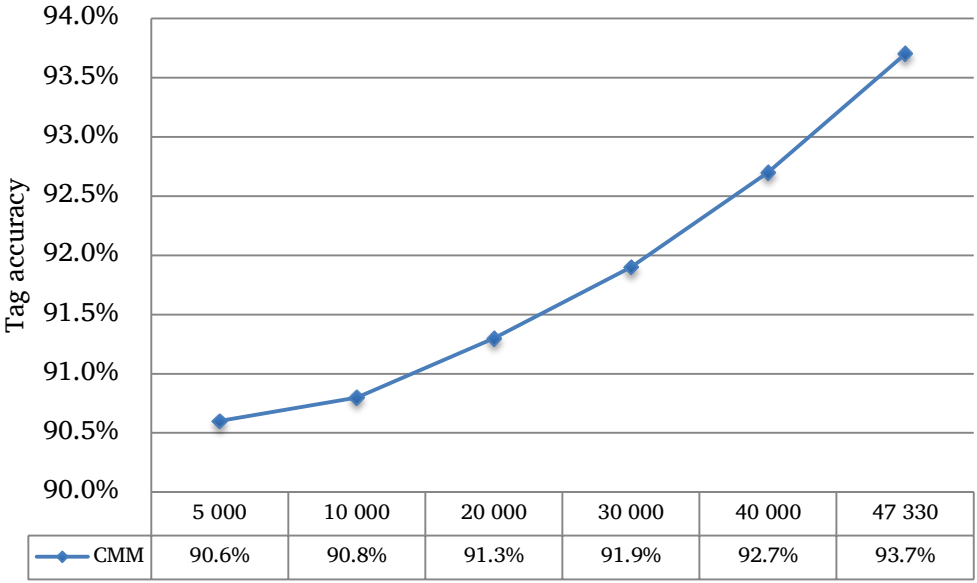


FIGURE 3 – Effect of corpus size on CMM disambiguation accuracy compared to single-token baseline

7 Conclusion and Outlook

We have developed a freely available morphological analysis and disambiguation solution for Latvian language. The tools, resources and corpora are publicly available under an open source licence.

We demonstrate that morphological analysis and disambiguation for languages with rich morphology can be performed with small amounts of language-specific resources. In particular, if the inflection rules can be formally defined, then a morphological tagging module with a useful accuracy of 90% can be trained even with a small annotated corpus of 10-20 thousand words and a limited dictionary.

We expect to further improve accuracy of the morphological tagger by extending the training data up to 100 000 words and exploring options for integration with syntactic parsers.

A future goal is to attempt to apply this methodology for Latgalian language – a regional language with approx. 165 000 native speakers and very limited digital resources. We also hope that this experience can inspire linguistic tool development for other languages with limited size of corpora, noting that a practically useful accuracy can be obtained with very limited language data.

Acknowledgments

We thank the University of Latvia for the financial support in preparing and presenting this paper, and the anonymous reviewers for their improvement recommendations.

References

Bārzdiņš G., Grūzītis N., Nešpore G. and Saulīte B. (2007). Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order. In *Proceedings of the 16th Nordic Conference of Computational Linguistics*, pages 13–20, Tartu.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004)*, pages 1535–1538, Paris.

Deksne, D. and Skadiņš, R. (2011). CFG Based Grammar Checker for Latvian. In *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011*, Riga, Latvia.

Drīzule, V. (1978). Об автоматическом распознавании омонимии флексий латышского языка [On automated recognition of flexive homonymy in Latvian language]. In *LZA Vēstis 1978*, 10, pages 79–87, Rīga, LZA.

Gahbiche-Braham, S., Bonneau-Maynard, H., Lavergne, T. and Yvon, F. (2012). Joint Segmentation and POS Tagging for Arabic Using a CRF-based Classifier. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Greitāne I. (1994). Latviešu valodas lokāmo vārdšķiru locīšanas algoritmi. (Algorithms for Latvian Form Generation) In *LZA Vēstis 1994*, 1, pages 32–39, Rīga, LZA.

Hajic, J., Krbec, P., Kveton, P., Oliva, K. and Petkevic, V. (2001). Serial combination of rules and statistics: A case study in Czech tagging. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 268–275.

Hajič, J. (2000). Morphological Tagging: Data vs. Dictionaries. In: *Proceedings of the 6th Applied Natural Language Processing and the 1st NAACL Conference*, pages 94-101, Seatle, Washington, U.S.A.

Hulden, M. and Francom, J. (2012). Boosting statistical tagger accuracy with simple rule-based grammars. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Levāne, K. and Spektors A. (2000). Morphemic Analysis and Morphological Tagging of Latvian Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, vol. 2, pages 1095–1098.

Levāne-Petrova K. (2011). Morfoloģiski marķēta valodas korpusa izmantošana valodas izpētē. In *"Vārds un tā pētīšanas aspekti": Rakstu krājums 15(1)*, pages 187–193, Liepāja, LiePA.

Paikens, P. (2007). Lexicon-based morphological analysis of Latvian language. In *Proceedings of 3rd Baltic Conference on Human Language Technologies (HLT 2007)*, pages 235-240, Kaunas.

Pinnis, M. and Goba, K. (2011). Maximum Entropy Model for Disambiguation of Rich Morphological Tags. In *Systems and Frameworks for Computational Morphology,*

Communications in Computer and Information Science, 1, Volume 100, The 2nd Workshop on Systems and Frameworks for Computational Morphology (SFCM2011), pages 14-22, Heidelberg, Springer.

Sarkans U. (1996). Morphemic and Morphological Analysis of the Latvian Language. In *Proceedings of the Forth conference on Computational Lexicography and Text Research*, pages 219–225, Budapest

Skadiņa I. (2004). Latviešu valodas morfoloģiskās analīzes sistēma – tās nozīme teikuma pareizrakstības pārbaudē. In *Vārds un tā pētīšanas aspekti 8*, pages 282–290, Liepāja.

Soida, E. and Kļaviņa, S. (1970). *Latviešu valodas inversā vārdnīca*, Rīga, LVU.

Toutanova K., Klein D., Manning C.D. and Singer Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, pages 252–259.

Vasiļjevs, A., Ķikāne, J. and Skadiņš, R. (2004). Development of HLT for Baltic languages in widely used applications. In *Proceedings of First Baltic Conference „Human Language Technologies – the Baltic Perspective"*, pages 198-202, Riga.

Yuret, D. and Türe F. (2006). Learning morphological disambiguation rules for Turkish. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL '06),* pages 328-334, Association for Computational Linguistics, Stroudsburg, PA, USA.