Morphological annotation of the Lithuanian corpus

Vidas Daudaravičius
Centre of Computational linguistics
Vytautas Magnus University
Donelaičio 58, Kaunas, Lithuania
vidas@donelaitis.vdu.lt

Erika Rimkutė
Centre of Computational linguistics
Vytautas Magnus University
Donelaičio 58, Kaunas, Lithuania
e.rimkute@hmf.vdu.lt

Andrius Utka
Centre of Computational linguistics
Vytautas Magnus University
Donelaičio 58, Kaunas, Lithuania
a.utka@hmf.vdu.lt

Abstract

As the development of information technologies makes progress, large morphologically annotated corpora become a necessity, as they are necessary for moving onto higher levels of language computerisation (e. g. automatic syntactic and semantic analysis, information extraction, machine translation). Research of morphological disambiguation and morphological annotation of the 100 million word Lithuanian corpus are presented in the article. Statistical methods have enabled to develop the automatic tool of morphological annotation for Lithuanian, with the disambiguation precision of 94%. Statistical data about the distribution of parts of speech, most frequent wordforms, and lemmas, in the annotated Corpus of The Contemporary Lithuanian Language is also presented.

1 Introduction

The goal of this paper is to present the experience and results of compiling a large Lithuanian morphologically annotated corpus by using an available Lithuanian morphological analyser and dealing with the disambiguation problem.

The Corpus of the Contemporary Lithuanian Language is a database of electronic texts, which is widely used in Lithuania. It well represents the present Lithuanian language and its different varieties (more about that in http://donelaitis.vdu.lt/).

<word="Nenuostabu" lemma="nenuostabus" type="bdvr neig nelygin.l neįvardž bevrd.gim">
<sep=",">
<word="kad" lemma="kad" type="jngt">
<word="muziejus" lemma="muziejus" type="dktv vyr.gim vnsk V">
<word="susilaukia" lemma="susilaukti(-ia,-ė)" type="vksm teig sngr tiesiog.nuos esam.l vnsk IIIasm">
<word="daugelio" lemma="daugelis" type="dktv vyr.gim vnsk K">
<word="svečių" lemma="svečias" type="dktv vyr.gim dgsk K">
<word="ne tik" lemma="ne tik" type="jngt">
<word="iš" lemma="iš" type="prln">
<word="Čikagos" lemma="Čikaga" type="tikr dktv mot.gim vnsk K">
<word="ir" lemma="ir" type="jngt">
<word="apylinkių" lemma="apylinkė" type="dktv mot.gim dgsk K">
<sep=",">
<word="bet ir" lemma="bet ir" type="jngt">
<word="tolimiausių" lemma="tolimas" type="bdvr teig aukšč.l neįvardž vyr.gim dgsk K">
<word="Amerikos" lemma="Amerika" type="tikr dktv mot.gim vnsk K">
<word="kampelių" lemma="kampelis" type="dktv vyr.gim dgsk K">
<word="bei" lemma="bei" type="jngt">
<word="kitų" lemma="kitas" type="įvrd mot.gim dgsk K">
<word="šalių" lemma="šalis" type="dktv mot.gim dgsk K">
<sep=".">

Figure 1: Extract from the morphologically annotated corpus (The following morphologically annotated sentence is presented: "It is no surprise that the museum is visited by guests not only from Chicago region, but also from distant American places and other countries.").

Balto-Slavonic Natural Language Processing 2007, June 29, 2007, pages 94–99,
Prague, June 2007.

Morphological annotation of the corpus will further increase capabilities of the corpus enabling extraction of unambiguous lexical and morphological information. The annotated corpus will soon be accessible for search on the internet. At the moment this corpus is fully accessible only at the Centre of Computational Linguistics of the Vytautas Magnus University. The tools for annotating Lithuanian texts are available for research purposes by request.

The Lithuanian morphological analyser *Lemuoklis* (Zinkevičius, 2000) produces results of morphological analysis of Lithuanian wordforms, but leaves unsolved the problem of morphological ambiguity. Considering successful application of statistical methods in solving the morphological ambiguity for other languages, statistical methods have also been chosen for Lithuanian. Research of morphological disambiguation and results of morphological annotation of the 100 million word Lithuanian corpus are presented in the article.

2 Morphological analysis of Lithuanian

Morphologically ambiguous wordforms are words or wordforms that have two or more possible lemma interpretations or morphological annotations, e. g. for the wordform *kovų* (en. *fights*, pl. Gen.) the morphological analyser *Lemuoklis* identifies two lemmas *kovas* (en. *rook* [bird] or *March* [month]) and *kova* (en. *fight*), while the wordform *naktis* (en. *night*) can be in Singular Nominative or in Plural Accusative case (more information on ambiguity for Lithuanian see Rimkutė, 2006).

Approximately a half of all wordforms in the Lithuanian annotated corpus are morphologically ambiguous (Rimkutė, 2006), which is comparable to other inflected languages, e.g. for the Czech language it is 46% (Hajič, 2004:173).

For developing the automatic disambiguation system a morphologically annotated training corpus is necessary. Manual creation of 1 M word Lithuanian annotated corpus is a very time consuming task, which has taken 5 man-years to complete. Firstly, the annotation format needs to be developed and mastered (see Figure 1), then it is necessary to assign a word to an appropriate part of speech, and often it is very difficult to find a correct grammatical reading for a word. It also takes a lot of time reviewing and trying to put all annotated texts into one uniform standard.

3 Automatic morphological annotation of the Lithuanian corpus

Statistical morphological disambiguation using small manually annotated training corpora looks as quite a simple task, when frequencies of grammatical features are generated during the training phase and the most likely sequence of morphological features is found in a new text by the help of various probability methods. Drawing on the experience of morphological annotation systems for other free word order languages (Dębowski, 2004; Hajič et al., 2001; Palanisamy et al., 2006 etc.), it is obvious that the corpus-based method is most suitable for the developing such systems for Lithuanian.

The Czech experience (Hladká, 2000) was very expedient for developing automatic morphological annotation tool for Lithuanian, especially because Czech similarly to Lithuanian is a free word order language. Czech research applies statistical Hidden Markov Models and formal rule-based methods for Czech and English languages. It is important to note that these methods are language independent and can be applied to Lithuanian. The only language dependent factor is a small morphologically annotated corpus for training. In various experiments the selection of Czech morphological features was regularized and optimised, which helped to achieve close to English language precision of 96%. However this precision is achieved with a limited number of Czech morphological features. The precision of 94 % is achieved when all features of Czech language are selected (Hladká, 2000).

4 Statistical morphological disambiguation

Morphologically analysed words are the input of the automatic morphological annotation system, while the best sequence of morphological features is its output. Annotation of a new text involves establishing the most likely sequence of morphological features by the help of Hidden Markov models. Not all combinations of trigrams and bigrams can be found even in the biggest corpora. Therefore, the linear smoothing of the missing cases is used, as the probability of the most likely sequence cannot be equal to zero (see more on HMM in Jurafsky (2000:305-307)).

The following HMM model is used by Czech scientists:

$$\Gamma \approx \max_T \tilde{p}(w_1 \mid t_{i_1}) * \tilde{p}(t_{i_1}) * \tilde{p}(t_{i_2} \mid t_{i_1}) * \\ * \prod_{t=3}^{n} \tilde{p}(w_t \mid t_{i_t}) * \\ * \tilde{p}(t_{i_t} \mid t_{i_{t-1}}, t_{i_{t-2}}), T = t_{i_1}, t_{i_2}, \ldots, t_{i_n}$$

We expanded the model by including the lemma. This procedure is important to Lithuanian, where different lemmas often have identical wordforms and morphological features. Therefore the probability of a lemma is also included:

$$\Gamma \approx \max_T \tilde{p}(w_1 \mid t_{i_1}) * \tilde{p}(w_1 \mid l_{i_1}) * \tilde{p}(t_{i_1}) * \\ * \tilde{p}(t_{i_2} \mid t_{i_1}) * \prod_{t=3}^{n} \tilde{p}(w_t \mid t_{i_t}) * \tilde{p}(w_t \mid l_{i_t}) * \\ * \tilde{p}(t_{i_t} \mid t_{i_{t-1}}, t_{i_{t-2}}), T = t_{i_1}, t_{i_2}, \ldots, t_{i_n}$$

where

$$\tilde{p}(w_t \mid t_{i_t}) = \lambda_w * p(w_t \mid t_{i_t}) + (1 - \lambda_w) * 1 / W_{t_{i_t}}$$

is the smoothed probability of a wordform and tag pair.

$$\tilde{p}(w_t \mid l_{i_t}) = \lambda_{w1} * p(w_t \mid l_{i_t}) + (1 - \lambda_{w1}) * 1 / L_{t_{i_t}}$$

is the smoothed probability of a wordform and lemma pair.

$$\tilde{p}(t_{i_t}) = \lambda_{01} * p(t_{i_t}) + (1 - \lambda_{01}) * 1 / C_T$$

is the smoothed probability of a tag.

$$\tilde{p}(t_{i_t} \mid t_{i_{t-1}}) = \lambda_{11} * p(t_{i_t} \mid t_{i_{t-1}}) + \\ + \lambda_{12} * p(t_{i_t}) + (1 - \lambda_{11} - \lambda_{12}) * 1 / C_T$$

is the smoothed probability of a bigram tag .

$$\tilde{p}(t_{i_t} \mid t_{i_{t-1}}, t_{i_{t-2}}) = \lambda_{21} * p(t_{i_t} \mid t_{i_{t-1}}, t_{i_{t-2}}) + \\ + \lambda_{22} * p(t_{i_t} \mid t_{i_{t-1}}) + \lambda_{23} * p(t_{i_t}) + \\ + (1 - \lambda_{21} - \lambda_{22} - \lambda_{23}) * 1 / C_T$$

is the smoothed probability of a trigram tag .

$$p(w_t \mid t_{i_t}) = \frac{Count\,(w_t \mid t_{i_t})}{Count\,(t_{i_t})}$$

is the probability of a wordform containing a particular tag in the training corpus.

$$p(t_{i_t}) = \frac{Count(t_{i_t})}{\mid T_{train} \mid}$$

is the probability of a tag in the training corpus.

$$p(t_{i_t} \mid t_{i_{t-1}}) = \frac{Count(t_{i_t}, t_{i_{t-1}})}{Count(t_{i_{t-1}})}$$

is the probability of a bigram tag in the training corpus.

$$p(t_{i_t} \mid t_{i_{t-1}}, t_{i_{t-2}}) = \frac{Count(t_{i_t}, t_{i_{t-1}}, t_{i_{t-2}})}{Count(t_{i_{t-1}}, t_{i_{t-2}})}$$

is the probability of a trigram tag in the training corpus.

$W_{t_{i_t}}$ is a number of wordforms with the feature t_{i_t}

$L_{t_{i_t}}$ is a number of lemmas with the feature t_{i_t}

C_T is a number of tags in T_{train} training set.

A function *Count(x)* corresponds to the frequency of a tag or a bigram.

Smoothing lambdas λ_{w1}, λ_w, λ_{01}, λ_{11}, λ_{12}, λ_{21}, λ_{22}, $\lambda_{23} < 1$ are used to combine the probabilities of lower order. The smoothing is very important when unknown events occur in the training corpus.

We used such lambda values:

λ_{w1} =0.85,

λ_w =0.85,

λ_{01} =0.99,

λ_{11} =0.74, λ_{12} =0.25,

λ_{21} =0.743, λ_{22} =0.203, λ_{23} =0.053

If a trigram tag is not found in the training corpus then the probability of a trigram is not assigned to zero, but rather the probability of a bigram is included with some weight. In case no trigram tag, bigram tag and unigram tag is found then the probability of a trigram assumes a very small number which is equal to 1 divided by the size of the tagset. The highest score is assigned to a trigram, lower – to bigram, and lowest – unigram.

The disambiguation tool has been developed at the Centre of Computational Linguistics of the Vytautas Magnus University using C++ tools. All results reported in this paper are based on approach using an accuracy criterion (number of correctly disambiguated results divided by number of input words). We do not use any morphological pre-processing. A precision of 94% has been achieved for establishing tags, which is comparable to results achieved for other languages, when the 1 million word training corpus is used. A precision of 99% is achieved for establishing lemmas. For the precision test a special 50 thousand word corpus has been used, which is not included in the training corpus.

The following statistics has been derived from the 1 M word training corpus[1]:

Different lemmas	*41,408*
Different pairs of wordforms and tags	*130,511*
Different pairs of wordforms and lemmas	*121,634*
Unigram tags C_T	*1,449*
Bigram tags	*76,312*
Trigram tags	*544,922*
Training corpus size $\vert T_{train} \vert$	*1,009,516*

Table 1: Corpus statistics

The number of lemmas in the training corpus is sufficient to gather frequencies in order to solve ambiguous lemmas. Unknown lemmas are not ambiguous in the training corpus, as they are rare and have unique meanings.

The size of the tagset is 1449. Lithuanian is a relatively free word order language, and therefore it is difficult to get reliable bigram and trigram statistics. We decided to gather distant bigram and trigram frequencies using a gap of 1. As a bigram we consider two subsequent tags (<A> <B>) or two tags with a gap of 1 in between (<A> <gap> <B>). Similarly, a trigram is a sequence of three subsequent tags (<A> <B> <C>) or a sequence of three tags with a gap of 1 between the first and second tag (<A> <gap> <B> <C>) or between the second and third tag (<A> <B> <gap> <C>). Distant n-grams help to reduce the number of unknown bigrams and trigrams in the training corpus.

5 Statistical data for the morphologically annotated corpus of Lithuanian

Most important statistical data for the morphologically annotated Lithuanian corpus:

- *Corpus size – 111,745,938 running words;*
- *Number of wordforms – 1,830,278;*
- *Number of unrecognized wordforms – 824,387 (5,6 % of all tokens);*
- *Number of recognized wordforms – 1,005,891.*

225,319 different lemmas have been recognized in the Corpus of Contemporary Lithuanian.

Distribution of parts of speech in the whole 100 M word corpus does not differ significantly from the distribution in the training corpus (see Figure 2). The biggest difference is in the number of unknown words. There are no unknown words in the training corpus, because it has been semi-automatically annotated and disambiguated. The number of unknown words in the 100 M word corpus is influenced by morphological analyzer, i.e., not all words are successfully recognized.

A big part of unknown words are proper nouns. Presently the dictionary of the morphological analyser contains 5255 high frequency proper noun lemmas (e.g. *Lietuva* (en. Lithuania)), which account for 3.2% of the vocabulary in the large annotated corpus. In the training corpus proper nouns account for 4.3% of the vocabulary, and we expect the similar proportion in the large annotated corpus. The average frequency of a proper noun lemma is 4.6 in the training corpus. Thus we could estimate the size of the dictionary of proper nouns at about 250,000 lemmas.

[1] See more about manually tagged Lithuanian Corpus and Lithuanian language tagset in Zinkevičius et al. 2005.

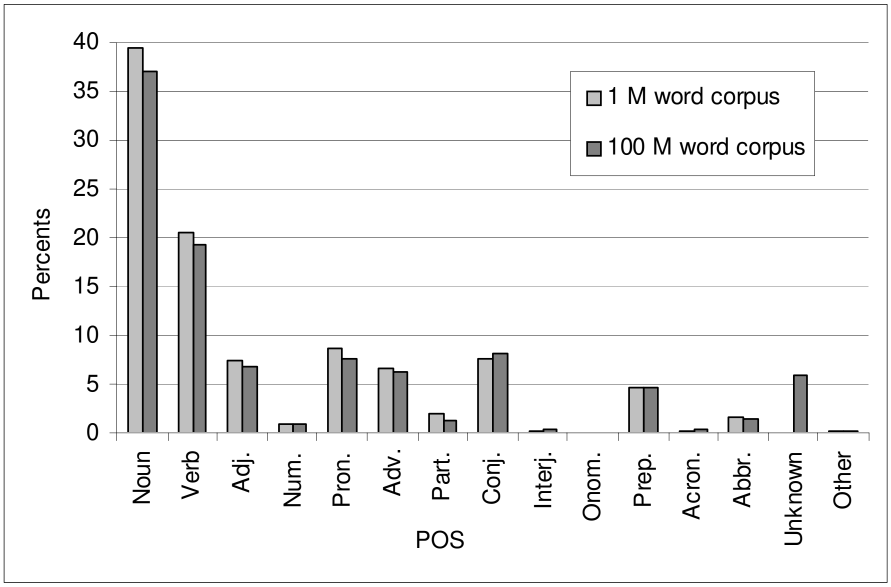

Figure 2: Distribution of parts of speech in 1 M and 100 M word corpora.

6 The remaining problems

The achieved precision of 94% for morphological annotation leaves some room for improvement. It is still difficult to solve homographic problems, where some wordforms of different words are identical. For example, wrong lemmas are frequently chosen for the wordforms *tonas* (en. *tone*) and *tona* (en. *ton*)*, kovas* (en. *rook* [bird]) and *kova* (en. *fight*)*, Biržai* (Lithuanian town) and *birža* (en. *stock-market*).

Syncretism of grammatical cases is not always solved correctly. Most often the incorrect analysis is given for words of feminine gender, when singular Genitive and plural Nominative cases are confused (e. g. *mokyklos* (en. *school*)).

Some cases are problematic even for a human linguist, when it is not clear which part of speech (noun or verb) is used in such collocations: *kovos dėl teisės likti pirmajame ešelone* (lit. *fight/ fights for the right to stay in the first league*); *kovos su narkotikais* (lit. *fight/ fights against drugs*); *kovos su okupantais* (lit. *fight/ fights against occupants*). Even if the part of speech of the word *kovos* is chosen as a noun, then the ambiguity case still remains. The broader context is needed to solve such problems.

Interjections are not very often used in Lithuanian, nevertheless the morphological abbreviation *a* is confused with the interjection *a.* Abbreviations that are identical to Roman numerals are often annotated incorrectly: the most problems are caused by the abbreviation *V.*

Sometimes wrong lemma is chosen. The words with fixed forms such as *ir* (en. *and*)*, tik* (en. *only*) cause many problems as they can be interjections, particles, or adverbs. The lemma of the wordform *vienas* (en. *one, alone, single)* is not always chosen correctly, as this word can be a pronoun, an adjective, a numeral, or even a proper noun. It is hoped that some of these problems will disappear after improving the program of morphological analysis.

7 Conclusions

The method of Hidden Markov models for morphological annotation has allowed achieving the precision of 94%, which is comparable to the precision achieved for other languages, when 1 M word training corpus is used. The precision of 99% is achieved for establishing lemmas of Lithuanian words. The precision measure estimates only the process of disambiguation, while unrecognised words are not included in the precision test.

The amount of unrecognised wordforms makes up 5,6% of all tokens (more that 800,000 different wordforms). In order to analyse the missing wordforms around 100-150 thousand lemmas need to be added to the lexicon of morphological analyser, i.e. the amount is similar to the present size of the lexicon.

One million word morphologically annotated corpus is enough for the analysis of morphological phenomena in Lithuanian, as distribution of parts of speech in the 100 million word corpus does not differ significantly

8 Acknowledgements

This work is a part of the project “Preservation of the Lithuanian Language under Conditions of Globalization: annotated corpus of the Lithuanian language (ALKA)”, which was financed by the Lithuanian State Science and Study Foundation.

References:

Arulmozhi Palanisamy and Sobha Lalitha Devi. 2006. HMM based POS Tagger for a Relatively Free Word Order Language. *Research in Computing Science 18*, pp. 37-48

Barbora Vidová-Hladká. 2000. Czech language tagging. Ph.D. thesis, ÚFAL MFF UK, Prague.

Daniel Jurafsky, James H. Martin. 2000. *Speech and Language Processing*, Prentice-Hall, Upper Saddle River, NJ.

Erika Rimkutė. 2006. Morfologinio daugiareikšmiškumo ribojimas kompiuteriniame tekstyne (Morphological Disambiguation of the Corpus of Lithuanian Language). Doctoral dissertation, Vytautas Magnus University, Kaunas.

Jan Hajič. 2004. *Disambiguation of rich inflection. Computational morphology of Czech*. Karolinum Charles University, Prague.

Jan Hajič, Pavel Krbec, Pavel Květoň, Karel Oliva, Vladimír Petkevič. 2001. Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. In *Proceedings of the 39Annual Meeting of the ACL (ACL-EACL 2001).* Université de Sciences Sociales, Toulouse, France.

Łukasz Dębowski. 2004. Trigram morphosyntactic tagger for Polish. In *Proceedings of the International IIS:IIPWM'04 Conference*, pp. 409-413, Zakopane.

Vytautas Zinkevičius. 2000. Lemuoklis – morfologinei analizei (A tool for morphological analysis - Lemuoklis). *Darbai ir Dienos*, 24, pp. 246–273. Vytautas Magnus University, Kaunas.

Vytautas Zinkevičius, Vidas Daudaravičius, and Erika Rimkutė. 2005. The Morphologically annotated Lithuanian Corpus. In *Proceedings of The Second Baltic Conference on Human Language Technologies*, pp. 365–370. Tallinn.

Appendix 1. Lithuanian morphological categories and appropriate tags

Grammatical Category	Equivalent in English	Tag
Abbreviation	dr.	sntrmp
Acronym	NATO	akronim
Adjective	good	bdvr
Adverb	perfectly	prvks
Onomatopoetic interjection	cock-a-doodle-do	ištk
Conjunction	and	jngt
Half participle	when speaking	psdlv
Infinitive	to be	bndr
Second Infinitive	at a run	būdn
Interjection	yahoo	jstk
Noun	a book	dktv
Number	one	sktv
Roman Number	I	rom skaič
Proper Noun	London	tikr dktv
Proper Noun2	Don	tikr dktv2
Participle	walking	dlv
Gerund	on the walk home	padlv
Preposition	on	prln
Pronoun	he	įvrd
Verb	do	vksm
Idiom AA	rest eternal	idAA
Connective idiom	et cetera	idJngt
P.S.	P.S.	idPS
Prepositional idiom	inter alia	idPrln
Pronominal idiom	nevertheless	idĮvrd
Particle	also	dll