THE MORPHOLOGICALLY ANNOTATED LITHUANIAN CORPUS

Vytautas Zinkevičius, Vidas Daudaravičius, Erika Rimkutė
Vytautas Magnus University (Kaunas, Lithuania)

Abstract

The paper deals with the preliminary findings from the morphologically annotated corpus of Lithuanian language (1 million running words). It was compiled and processed at the Center of Computational Linguistics, Vytautas Magnus University. Each annotation for an inflected word form of the corpus contains a lemma and a set of morphological features. The paper presents the strategy for automatic and manual annotation. Automatic annotation was carried out with the help of analyser-lemmatiser. Disambiguation of the homoforms was performed manually. Tag sets and the most prominent features of Lithuanian morphology are discussed in detail. The annotated corpus allowed us to measure the usage of parts of speech and their morphological features in contemporary Lithuanian language. The annotated corpus is of great importance for future development of parsing tools, treebanks and other NLP tools and resources for Lithuanian language.

Keywords: corpus compilation, morphological annotation, Lithuanian language, tag sets, morphological disambiguation

1. Introduction

The paper describes processing of the first Lithuanian annotated corpus (LAC) at the Center of Computational Linguistics (CCL), Vytautas Magnus University. LAC is important since it enables linguistic, statistical and computer science research. Moreover, it gives the ground for future development of language technologies for Lithuanian, e.g. automatic morphological disambiguators, parsing tools, treebanks. Texts for compiling LAC were selected from the Corpus of the Contemporary Lithuanian Language at CCL (more about see Marcinkevičienė et al. 2004), which comprises more than 100 million running words.

LAC is a set of XML files, containing 1 million running words annotated morphologically. Each annotation for a word form contains its normalized form (lemma), and a full set of morphological properties for the inflected word form. Non-word textual units, such as punctuation marks, spaces, paragraphs, numbers, are represented in LAC by special marks.

The annotation of the LAC started at 2001. In the beginning the collection of the texts was compiled. The collection had to present a wide range of textual genres and registers in order to have as much of the variety of morphological information as possible, e.g. scientific texts, fiction, parliament debates and administrative texts, to mention a few.

Table 1. The set of POS tags and their abbreviations

Grammatical Category	Equivalent in English	Tag
Abbreviation	dr.	Abbr
Acronym	NATO	Acronym
Adjective	good	Adj
Adverb	perfectly	Adv
Onomatopoetic interjection	cock-a-doodle-do	Onom
Conjunction	and	Conn
Half participle	when speaking	Half_part
Infinitive	to be	Inf
Second Infinitive	at a run	Inf2
Interjection	yahoo	Interjection
Noun	a book	N
Number	one	Numb
Roman Number	I	Numb2
Proper Noun	London	PN
Proper Noun2	Don	PN2
Participle	walking	Part
Gerund	on the walk home	Gerund
Preposition	on	Prep
Pronoun	he	Pron
Verb	do	V
Idiom AA	rest eternal	idAA
Connective idiom	et cetera	idConn
P.S.	P.S.	idPS
Prepositional idiom	inter alia	idPrep
Pronominal idiom	nevertheless	idPron
Particle	also	Particle

Table 2. The set of the morphological features and their abbreviations

Property	Value	Tag
Reflexiveness	reflexive	Ref
	non-reflexive	NonRef
Positiveness	positive	Pos
	negative	Neg
Voice	active voice	ActVoice
	passive voice	PassVoice
	participle of necessity	PartOfNecess
Mood	indicative mood	IndicatM
	imperative mood	ImperatM
	subjunctive mood	SubjunctM
Tense	present tense	PresT
	simple past tense	SimplePastT
	past frequentative tense	PastFreqT
	past tense	PastT
	future tense	FutT
Numeral type	cardinal numeral	CardinalNumb
	plural numeral	PlurNumb
	collective numeral	CollectNumb
	ordinal numeral	OrdinalNumb
Degree	positive degree	PosDeg
	comparative degree	CompDeg
	attenuated degree	AttenDeg
	superlative degree	SupDeg
Definiteness	definitive	Def
	undefinitive	Undef
Gender	common gender	Comm
	masculine gender	Masc
	feminine gender	Fem
	neuter gender	Neut
Number	singular	Sg
	plural	Pl
	dual	Dual
Case	nominative	Nom
	genitive	Gen
	dative	Dat
	accusative	Acc
	instrumental	Inst
	locative	Loc
	vocative	Voc
	illative	Il
Person	first person	I
	second person	II
	third person	III

LAC is the first Lithuanian morphologically annotated corpus that includes full texts of various genres and registers. Nevertheless a few databases containing data about lexical and inflectional frequencies could be mentioned here as being closest products for Lithuanian language (Grumadienė 2002; Mauricaitė et al. 2004). In comparison with annotated corpora for other languages LAC is of similar size as many other manually annotated corpora but different in annotation level. For instance, the *TIGER* corpus of German has 1 million words (Brants et al. 2002), the treebank of Russian has 1 million words (Boguslavsky et al. 2000), the treebank of Czech consists of 1,5 million words (Hajič 2002). These above-mentioned corpora are annotated morphologically and syntactically. Prague Dependency Treebank is also annotated tectogrammatically (Hajičová 1998). Tag set of

each corpus is different in size and in features. Mostly it depends on the type of the language, i.e. inflected languages have a richer tags sets.

2. Corpus Annotation

The Lithuanian language is a highly inflected language, e.g. ending *-o* is grammatically polysemous since it denotes singular Genitive of masculine noun, pronoun or numeral, e.g. *šito vieno aukšto perstatyto pastato (of this one floor rebuilt house)*. This ending can also be the third person of present or past tense verb form, e.g *daro* (*he does/they do*), *ėjo* (*he/they went*). Tables 1 and 2 show the system of tags used for the morphological annotations in LAC.

The example from LAC shows the structure of the tags and text annotation. "word" presents the word form or token used in the text, "lemma" is the normalized form of the word form and "type" is the set of morphological features which describes the word form.

The example from LAC:

```
<word="Skulptūra" lemma="skulptūra" type="N Fem Sg Nom">
<space>
<word="buvo" lemma="būti(yra,buvo)" type="V Pos NonRef ActVoice SimplePastT Sg III">
```

Every part of speech has its specific set of tags, e.g. noun is usually described with a help of three features: gender, number and case. However some loan words lack some features, e.g. the word "taxi" in Lithuanian is not inflected and it is difficult to assign gender, number and case. In this case no features, only part of speech tag is assigned to "taxi", i.e., <word="taxi" lemma="taxi" type="N">. Some morphological multiword units that can not be analyzed separately required specific additional tags such as pronominal idioms, prepositional idioms, connection idioms. Besides one additional case tag for the obsolete illative case missing in contemporary Lithuanian language was included.

Non-lexical units were marked using separate tags, e.g., spaces were marked as "<space>", all punctuation marks were put in the tag <sep="…"> (e.g. <sep="!">), numbers were put in the tag <number="…"> (e.g. <number="86">), foreign insertions were marked between "<foreign lang="…">" (e.g. <foreign lang="en">) and "</foreign>".

The process of LAC annotation is presented in Figure 1. Possible lemmas and the morphological features were automatically assigned to all word forms using the morphological analyzer-lemmatizer Lemuoklis (Zinkevičius 2000). Since the tool processes only isolated words, it produces all theoretically possible grammatical interpretations for each inflected word form of text, including all possible lemmas. Not all word forms were recognized therefore manual assignment of lemmas and morphological features was necessary. Lemmatization was followed by manual disambiguation since almost half of the word forms were ambiguous in the morphological sense.

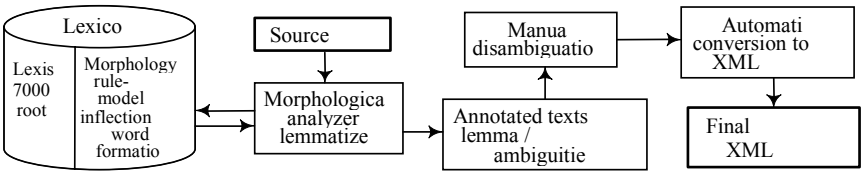

Figure 1. The process of LAC annotation

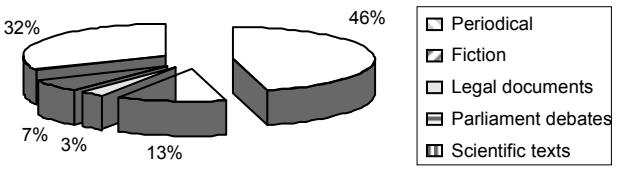

Figure 2. The textual structure of LAC

3. Morphological ambiguity and the resolution

Manual disambiguation revealed that circa 47 % of Lithuanian words or word forms are morphologically ambiguous. The morphological ambiguity of inflected languages is similar, e.g. the morphological ambiguity of Czech language is ca 46 % (Hajič 2004: 173). There are two types of morphologically ambiguous words or word forms: lemma ambiguity and word form ambiguity. An example of lemma ambiguity is *namo* – noun singular Genitive (*namas*) and adverb (*namo*). An example of word form ambiguity is *mamos* (*mother's* or *mothers*) – singular Genitive or plural Nominative.

Lemuoklis can not recognize ca 11 % of word forms of the corpus. The most typical unidentified cases are: proper nouns, foreign words, abbreviations, acronyms, abbreviated forms, etc. Only ca 40 % of all word forms are morphologically unambiguous, e.g. *aš (me/I)* always has one lemma and one morphological tag (Rimkutė 2003: 60–78).

4. The structure of the LAC and the grammatical specificity of the Lithuanian language

The LAC contains ca 1 013 000 running words, ca 145 000 types (word forms), ca 49 000 lemmas. The structure of the LAC is presented in Figure 2. The average number of the word form/lemma ratio is 3.6. This means that one lemma has a bit more then three word forms in average. This ratio for Lithuanian language is much higher in comparison with the same for Polish that is equal 2.01 (the ratio was calculated on IPI PAN corpus (Przepiórkowsky 2004)). This means that Lithuanian word forms are much more inflected than Polish, despite the fact that not all parts of speech are inflected (see Figure 3). Only six parts of speech out of eleven are inflected: noun, verb, pronoun, adjective, participle and number. The most inflected part of speech is pronoun which lemma has up to seven word forms in average. Since pronouns cover a considerable part of the usage of contemporary Lithuanian (see Figure 4 for details) this means that pronouns are one of the most difficult parts of speech for grammatical analysis.

The largest amount of usage of Lithuanian language is dominated by nouns i.e. more than 36 %. Finite and non-finite

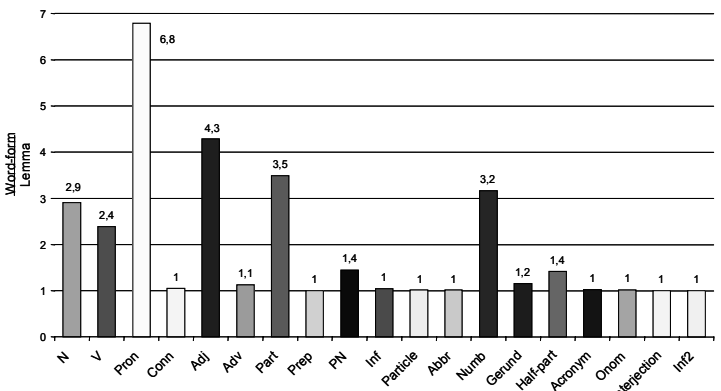

Figure 3. Word form/lemma ratio for part of speech

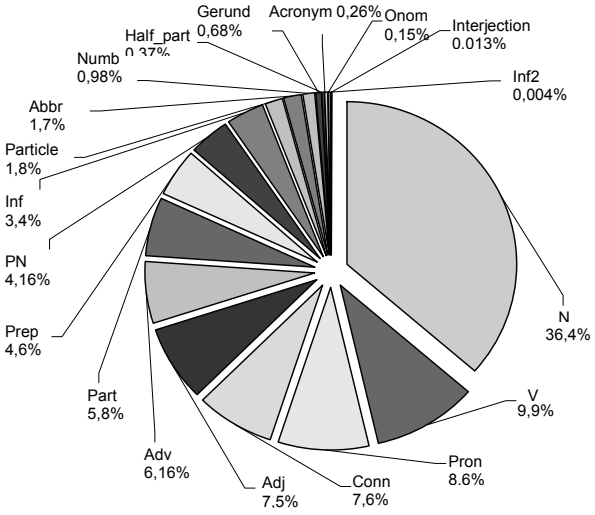

Figure 4. Distribution of parts of speech in LAC

verbs (infinitive, participle, half-participle, gerund, second infinitive) make up 20 % of the usage of Lithuanian language.

5. Concluding remarks

Annotation of the first Lithuanian language corpus provided with some valuable data concerning the grammatical specificity of the Lithuanian language. It turned out that inflection in real usage is not so prominent as in the grammatical system since highly inflected parts of speech such as verbs and nouns have less than 3 word-forms in average. Pronoun demonstrated surprisingly big number of word forms actually used in contemporary Lithuanian language. Overall, the tendencies for usage of different parts of speech coincide with the data obtained by other researchers as well as native speaker's intuitions, i.e. noun has the biggest coverage but verb is also very important. Thus Lithuanian language preserves its verbal nature.

The next step for processing of LAC will be its syntactic annotation with a help of Dependency Grammar formalism. LAC will be also used as a machine learning basis for the annotation of 100 million word corpus. Finally, LAC will be available on the web for research purposes.

Acknowledgement

We are grateful to the State Commission of the Lithuanian Language for supporting the compilation of the LAC in 2001–2003.

References

Boguslavsky, I.; Grigorieva, S.; Grigoriev, N.; Kreidlin, L.; Frid, N. 2000. Dependency treebank for Russian: concept, tools, types of information. In: *Proceedings of 18th International Conference on Computational Linguistics (COLING-2000).* Saarbrüken, Germany. Vol 2, 987–991.

Brants, Sabine; Dipper, Stefanie; Hansen, Silvia; Lezius, Wolfgang; Smith, George 2002. The TIGER treebank. In: *Proceedings of the Workshop on Treebanks and Linguistic Theories*. Sozopol, Bulgaria. Retrieved February 15, 2005, from http://www.coli.uni-sb.de/~sabine/tigertreebank.pdf.

Grumadienė, Laima 2002. Dabartinės rašomosios lietuvių kalbos dažninis žodynas ir jo bazė [The frequency dictionary of contemporary written Lithuanian and its data base]. In: Holvoet, A. (ed.) *Acta Linguistica Lithuanica* 46. Vilnius: Lietuvių kalbos institutas (Institute of the Lithuanian Language). 19–37.

Hajič, Jan 2002. Tectogrammatical representation: towards a minimal transfer in machine translation. In: *Proceedings of the Sixth International Workshop on*

Tree Adjoining Grammar and Related Frameworks (TAG+6). Venezia, Italy: Universita di Venezia. 216–226.

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

Hajičová E. 1998. Prague Dependency Treebank: from analytic to tectogrammatical annotations. In: Sojka, Petr; Matoušek, Václav; Pala, Karel; Kopeček, Ivan (eds.) *Text, Speech, Dialogue*. Brno: Masarykova univerzita. 45–50.

Marcinkevičienė, R.; Bielinskienė, A.; Daudaravičius, V.; Rimkutė, E. 2004. Corpora for Lithuanian language technologies. In: *The First Baltic Conference. Human Language Technologies. The Baltic Perspective*. Riga, Latvia, April 21–22, 2004. 21–24.

Mauricaitė, Vera; Norkaitienė, Milda; Pakerys, Antanas; Petrokienė Ritutė 2004 (eds.). Bendriniai XX a. spaudos žodžiai. Elektroninis dažninis žodynas [Common press words of the 20th century. Electronic frequency dictionary]. Vilnius: MELI [Science & Encyclopaedia Publishing Institute].

Przepiórkowski, Adam 2004. The IPI PAN corpus: preliminary version. Warszawa: Instytut podstav infromatyki PAN. Retrieved February 17, 2005, from http://dach.ipipan.waw.pl/~adamp/Papers/2004-corpus/book_en.pdf.

Rimkutė, Erika 2003. Morfologinio daugiareikšmiškumo tipologija [The typology of morphological ambiguity]. In: Merkys, V.; Ambrazas, V.; Sauka, L. (eds.) *Lituanistica* 4 (56). 60–78.

Zinkevičius, Vytautas 2000. Lemuoklis - morfologinei analizei [Morphological analysis with Lemuoklis]. In: Gudaitis, L. (ed.) *Darbai ir Dienos* 24. 246–273.

VYTAUTAS ZINKEVIČIUS is engineer programmer at the Center of Computational Linguistics, Vytautas Magnus University. He has graduated from Vilnius University at 1981. His research interests concern natural language resources, theory and tools for language analysis and generation, computational morphology for highly inflected languages, computer readable lexicons. His doctoral study focuses on digital modeling of the Lithuanian morphology. He is a developer of computational models for morphology and lexis that were implemented in several Lithuanian spell-checkers, and in compiling frequency dictionaries for contemporary Lithuanian. He also participates in creating the digital version of Dictionary of Lithuanian in 20 vol., an investment project at the Institute of Lithuanian Language. E-mail: vytasz@lki.lt

VIDAS DAUDARAVIČIUS is senior engineer-programmer and early-stage researcher at the Centre of Computational Linguistics Vytautas Magnus University. He received his M.A. (Applied Informatics) at VMU Faculty of Computer Science. His fields of research: computational syntax and parsing, computational linguistics, information retrieval and extraction, machine translation.

ERIKA RIMKUTĖ is a junior researcher of the Centre of Computational Linguistics at Vytautas Magnus University. She received her M. A. (Lithuanian language) at VMU. Her research interests include corpus linguistics, computational linguistics, automatic morphological analysis and synthesis, morphological ambiguity and disambiguation and automatic syntactic analysis. Her doctoral study focuses on morphological ambiguity and disambiguation in the Lithuanian language. E-mail: e.rimkute@hmf.vdu.lt.