A Comparison of Lithuanian Morphological Analyzers

Jurgita Kapočiūtė-Dzikienė[1(✉)], Erika Rimkutė[2], and Loic Boizou[2]

[1] Department of Applied Informatics, Vytautas Magnus University,
Vileikos str. 8, 44404 Kaunas, Lithuania
jurgita.kapociute-dzikiene@vdu.lt

[2] Centre of Computational Linguistics, Vytautas Magnus University,
Putvinskio str. 23, 44243 Kaunas, Lithuania
{erika.rimkute,loic.boizou}@vdu.lt
http://if.vdu.lt/
http://tekstynas.vdu.lt/

Abstract. In this paper we present the comparative research work disclosing strengths and weaknesses of two the most popular and publicly available Lithuanian morphological analyzers, in particular, *Lemuoklis* and *Semantika.lt*. Their lemmatization, part-of-speech tagging, and fined-grained annotation of the morphological categories (as case, gender, tense, etc.) performance was evaluated on the morphologically annotated gold standard corpus composed of four domains, in particular, administrative, fiction, scientific and periodical texts. *Semantika.lt* significantly outperformed *Lemuoklis* by ∼1.7%, ∼2.5%, and ∼8.1% on the lemmatization, part-of-speech tagging, and fine-grained annotation tasks achieving ∼98.0%, ∼95.3% and, ∼86.8% of the accuracy, respectively.

Semantika.lt was also superior on the administrative, fiction, and periodical texts; however, *Lemuoklis* yielded similar performance on the scientific texts and even bypassed *Semantika.lt* in the fine-grained annotation task.

Keywords: Lithuanian morphological analysers · Gold-standard corpus · Experimental evaluation · The Lithuanian language

1 Introduction and Related Work

If excluding so-called isolating languages as Mandarin Chinese which do not have grammatical categories (as case, gender, number, tense, etc.), languages show a varied degree of inflection (and derivation), starting from weakly inflected as English and going to highly inflected as Spanish, Czech, Lithuanian, Turkish, Arabic or Hebrew. In this paper we focus on the fusional Lithuanian language, which has the rich inflectional morphology even complex to Latvian or Slavic languages [20]. Different morphological categories are defined with various endings attached to the stable parts of words (i.e., to a root or to a root with affixes). In highly inflectional languages hundreds of word's forms can be generated from

K. Ekštein and V. Matoušek (Eds.): TSD 2017, LNAI 10415, pp. 47–56, 2017.
DOI: 10.1007/978-3-319-64206-2_6

a single root (e.g., ~2–3 million grammatical forms can be used for a dictionary of ~100 thousand words [9]); moreover, these forms often match other grammatical categories or parts-of-speech. Thus, a rate of ambiguous morphological forms for the Lithuanian language reaches even ~47 [18].

Morphological analysis has experienced a great success since the invention of the Two-Level morphology and the development of the finite-state technology that Two-level formalism is based on [14]. All existing morphological analyzers according to their creation method can be divided into knowledge-based (sometimes called rule-based and/or lexicon-based), supervised, and unsupervised. Despite that unsupervised approaches (segmenting the raw text into morphs as, e.g., *un+fail+ing+ly*) have become very attractive recently (because do not require gold morphological labels and for any language there is an unlimited number of text resources), they, however, are more suitable for agglutinative languages [3]. Knowledge-based approaches rely on rules/lexicons prepared by linguist-experts and do not require additional resources. Probably due to this reasons, this approach is still the most widely spread, thus, used for many different languages: English [11,19], French [7], Russian and Spanish [14], Urdu and Hindi [1,5], Tamil [2], etc.

Corpus-based morphological analyzers are the closest alternative to knowledge-based approaches. Although such systems are already built automatically in the supervised manner, induced rules are based on gold morphological annotations found in the training data. The annotation process itself is very laborious and requires deep language expertise, but such analyzers can be easily redeveloped and improved after adding more annotated texts. Analyzers of this type are used for many languages: Dutch [6], Swahili [17], Hindi [15], Kazakh [12], Arabic [13], Polish [10], etc.

Morphological analysis is important in such NLP applications as information retrieval, parsing or machine translation (especially when translating direction points from/to the morphologically rich language). Each module of such complex system has to be as accurate and reliable as possible (because the overall accuracy depends on cumulative accuracies of separate modules), including the morphological analyzer. The priority is its accuracy, no matter if the analyzer is developed using rule-based or corpus-based approach. The aim of this research is to evaluate, to compare and to determine the most accurate morphological analyzer for the Lithuanian language.

2 The Lithuanian Morphological Analyzers

The Lithuanian language is spoken by only ~3.2 million people world-wide; therefore it is not very attractive for big companies. Nevertheless this field of research has a rather long history. The first prototype of the Lithuanian morphological analyzer was created ~30 years ago and ever since there were several attempts towards creation of the accurate tool coping with the complex Lithuanian morphology. Despite all of those attempts, there are only two reliable morphological analyzers and lemmatizers which are still maintained, updated, and publicly available on-line:

1. *Lemuoklis*[1] at the beginning was purely rule and lexicon-based approach (described in detail by it's founder V. Zinkevičius in [22]), later extended with the statistical approach for the disambiguation of morphological homoforms. In *Lemuoklis* the knowledge about the Lithuanian language is stored in the lexical and grammar database, which contains 6 lexicons with various Lithuanian lexical groups; proper nouns, in the forms they as found in the corpus (that is without lemmatization); the stems of these proper nouns; obsolete and dialectal word forms; forms with the shortened endings which appear in literary and colloquial styles; abbreviations and acronyms. Since the database also contains word stems, each stem can be augmented with the affixes (prefixes, suffixes, endings) which, in turn, are determined according to the word's morphological type (i.e., its morphemic structure). Each analyzed word is divided on the basis of the various scheme options using prefix + stem + postfix pattern, therefore the implemented inflectional models not only recognize different inflectional word forms (including obsolete or dialectal as e.g., illative) of the existent words, but also synthesize some derivatives in their various inflected forms. The lexical database contains ∼91 thousand different headwords in total; however, the number of theoretically possible grammatical word forms can reach even several billions.

   Lemuoklis has been used for many practical tasks. One of its first versions was used in preparing the first frequency world list for the Lithuanian language [21], and it was later integrated into the Microsoft Office package and Information Base components and used for the automatic spell checking [22]. In 2000–2005 *Lemuoklis* was applied on ∼1 million word corpus, which afterwards was manually corrected by a linguist-expert (for more information see [18]) and led to the creation of the first lemmatized and morphologically annotated gold-standard corpus for the Lithuanian language. This research showed that within the ∼89% of all automatically recognized Lithuanian words no less than ∼47% are ambiguous. The disambiguation problem was solved out by complementing the rule and lexicon based approach with the statistical trigram Hidden Markov Model method (described in [8]): this version reached ∼94% of accuracy for annotation and ∼99% for lemma assignment.
2. *Semantika.lt*[2] morphological analyzer was created with the ambition to outperform its ancestor *Lemuoklis*, which still has not got rid of such shortcomings as rather low performance on the proper nouns. The main reason for designing a new tool, was the fact that *Lemuoklis* data was hard coded, which makes difficult to enrich the lexical database. *Semantika.lt* is also based on the hybrid approach: it is based on the Hunspell open source platform (consisting of the lexicon and the affixes) supplemented with the statistical method for the disambiguation task. The information included into the lexicon was taken from the following sources: from the 6th edition of the *Modern*

[1] At http://tekstynas.vdu.lt/page.xhtml;jsessionid=C27B0743101187E540CD32D0498C9887?id=morphological-annotator.

[2] At http://www.semantika.lt/TextAnnotation/Annotation/Annotate.

Lithuanian Dictionary; from the *Corpus of the Contemporary Lithuanian Language* at the Centre of Computational Linguistics of Vytautas Magnus University (∼100 million tokens; ∼600 thousand unique); from the database of the Lithuanian Parliamentary documents (∼400 million tokens; ∼1 million unique); from various public Internet sources. The created analyser resulted in 429 groups of rules; 1,518 explicit tags for flexing/non-flexing properties; 5,832 rules for suffix and affix alternation in 16,734 alternation cases. The total number of headwords in *Semantika.lt* is ∼146 thousand of which ∼38 thousand are common nouns, ∼67 thousand proper nouns, ∼12 thousand adjectives, ∼23 thousand verbs, ∼4 thousand words from other classes. The disambiguation problem as in *Lemuoklis* is solved using statistical trigram Markov model + Viterbi algorithm.

Thus, according to the number of headwords, *Semantika.lt* obviously outperforms *Lemuoklis*; but on the other side, *Lemuoklis* uses the synthesis method in order to handle some frequent derivation patterns (e.g. some regular agentive and diminutive forms). Besides, *Lemuoklis* had been updated in the past, but since 2007 it has not been experimentally evaluated. Moreover, *Semantika.lt* has never been fully evaluated on the basis of a gold-standard corpus. In general, there was no evaluation with explicit methodology. Therefore currently it is not clear which one is more accurate and whether difference in their accuracy is statistically significant.

The contribution of this research is to evaluate both of these analyzers and to compare their results following standard up-to-date methods for tool evaluation. However, that the research would be carried out correctly it is important (1) to equalize experimental conditions for the both analyzers (to evaluate them on the same gold-standard corpus; to equalize their annotation tags; to use the same evaluation metrics); (2) to test them on the unseen corpus which was neither used in the rule or lexicon creation nor in training for the disambiguation problem solving. Besides, we anticipate that the publicly available morphologically annotated gold-standard corpus (presented in this paper) could be treated as the benchmark corpus and used for evaluation and comparison purposes of other existing or forthcoming morphological analyzers.

3 The Comparative Evaluation

The experimental comparison (described in Sect. 3.2) of both Lithuanian morphological analyzers (presented in Sect. 2) was performed on a morphologically annotated gold-standard corpus (described in Sect. 3.1). The issue of the annotation format discordance (in the gold corpus and texts produced by both analyzers) was solved out by converting all formats to one based on the Leipzig glossing rules [4] used in the Universal Dependencies Project[3].

[3] More about the Universal Dependencies Project is presented in http://universaldependencies.org/.

3.1 The Gold-Standard Corpus

The first morphologically annotated gold-standard corpus (called MATAS[4]) was prepared by the Centre of Computational Linguistics at Vytautas Magnus University. It contains 1,641,263 words and covers 4 domains, in particular, administrative, fiction, scientific and periodical texts. MATAS was prepared in a semi-automatic manner: the initial annotations were obtained with *Lemuoklis* and afterwards manually verified and corrected by one linguist-expert.

Unfortunately for our experiments we could not take the entire corpus, because some parts of it have already been used in training of the *Semantika.lt* morphological analyzer. Thus, we had to select and annotate additional texts taking into account two important factors: (1) they must not have been used in creation/training of *Lemuoklis* and *Semantika.lt*; (2) the obtained gold-standard corpus has to be balanced (in terms of words) that results would not be biased towards the largest domains. Hence, for experiments we selected texts that contain ∼5 thousand words in each domain, resulting ∼20 thousand in totals. The statistics about the gold-standard corpus is presented in Table 1.

Table 1. The distribution of total and distinct (in brackets) words over different parts-of-speech and domains in the gold-standard corpus. The *unrecognized words* caption defines foreign language or misspelled Lithuanian words.

Part-of-speech	Administrative	Fiction	Scientific	Periodicals	All domains
Noun	2,102 (911)	1,305 (1,034)	2,158 (1,092)	1,768 (1,105)	7,333 (3,492)
Verb (all forms)	834 (513)	1,098 (869)	773 (506)	1,045 (754)	3,750 (2,306)
Adjective	469 (317)	372 (334)	557 (416)	313 (279)	1,711 (1,234)
Conjunction	364 (14)	497 (26)	355 (22)	340 (24)	1,556 (33)
Pronoun	186 (80)	754 (193)	179 (84)	398 (140)	1,517 (273)
Adverb	159 (74)	411 (175)	153 (85)	252 (120)	975 (302)
Proper noun	151 (34)	29 (27)	362 (190)	422 (239)	964 (462)
Preposition	155(15)	250 (25)	135 (15)	253 (25)	793 (31)
Particle	50 (16)	333 (48)	36 (9)	151 (34)	570 (64)
Numeral	179 (114)	25 (19)	165 (96)	137 (91)	506 (239)
Unrecognized word	89 (22)	11 (10)	105 (37)	114 (38)	319 (94)
Interjection	0 (0)	3 (3)	0 (0)	2 (2)	5 (4)
In total	4,738 (2,110)	5,088 (2,763)	4,978 (2,552)	5,195 (2,851)	19,999 (8,534)

3.2 The Experimental Set-Up and Results

In our experiments we compared the gold annotations with the automatic annotations produced by *Lemuoklis* and *Semantika.lt* and calculated the *accuracy* and *f-score* values.

Moreover, we evaluated if the differences between the results obtained by different morphological analyzers are statistically significant. The evaluation was

[4] The annotated corpus can be downloaded from https://clarin.vdu.lt/xmlui/handle/20.500.11821/9.

done using the McNemar test [16] with one degree of freedom at the significance level of $\alpha = 0.05$, meaning that the differences are considered statistically significant if calculated probability density function $p < \alpha$.

The obtained lemmatization accuracies are presented in Fig. 1. The *micro-accuracy* (or *micro-f-score*) values for the parts-of-speech (i.e., coarse-grained morphological information) and the morphological categories (i.e., fine-grained information as case, gender, number, voice, tense, etc.) are summarized in Fig. 2. The *f-score* values distributed over the different parts-of-speech are presented in Table 2.

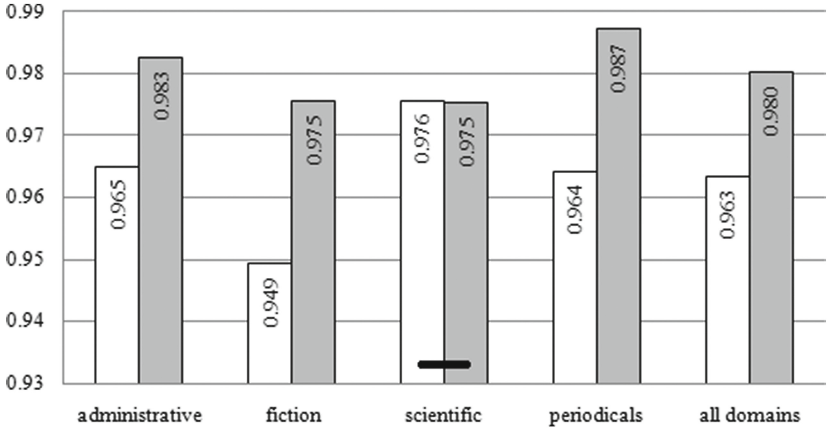

Fig. 1. The lemmatization *accuracies* in white and gray columns for *Lemuoklis* and *Semantika.lt*, respectively. The results which differences are not statistically significant are connected with a solid black line (see the *scientific* domain).

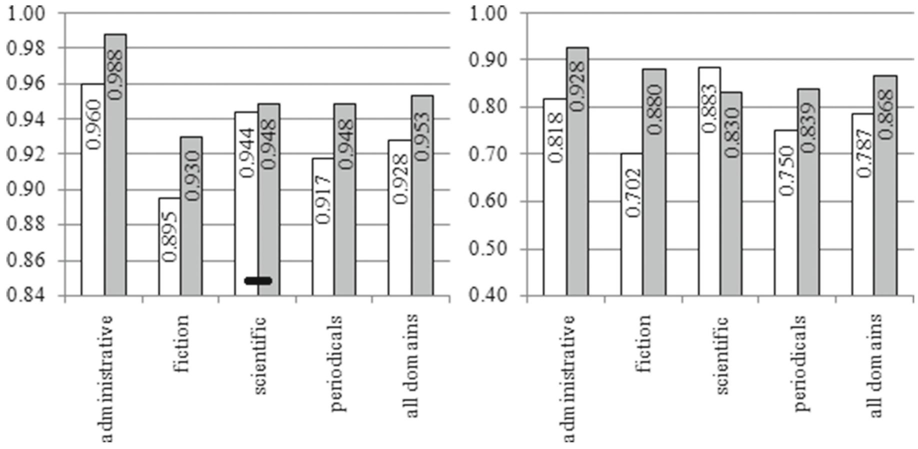

Fig. 2. The *micro-accuracy/micro-f-score* values for parts-of-speech (the left diagram) and morphological categories (the right diagram) in white and gray columns for *Lemuoklis* and *Semantika.lt*, respectively.

Table 2. The calculated *f-score* values for various parts-of-speech in different domains. *Lem* and *Sem* stands for *Lemuoklis* and *Semantika.lt*, respectively.

Part-of-speech	Administr.		Fiction		Scientific		Periodicals		All domains	
	Lem	Sem	Lem	Sem	Lem	Sem	Lem	Sem	Lem	Sem
Noun	0.995	0.997	0.976	0.983	0.989	0.993	0.983	0.989	0.987	0.992
Verb (all forms)	0.974	0.992	0.966	0.976	0.993	0.971	0.987	0.992	0.979	0.983
Adjective	0.955	0.982	0.919	0.948	0.986	0.953	0.939	0.944	0.954	0.958
Conjunction	0.957	0.992	0.823	0.893	0.966	0.948	0.872	0.923	0.896	0.934
Pronoun	0.943	0.984	0.907	0.964	0.977	0.969	0.908	0.959	0.920	0.966
Adverb	0.862	0.953	0.806	0.860	0.923	0.872	0.825	0.868	0.837	0.879
Proper noun	0.873	0.969	0.462	0.711	0.706	0.889	0.754	0.902	0.750	0.903
Preposition	0.926	0.997	0.982	0.982	0.989	1.000	0.986	0.992	0.973	0.991
Particle	0.472	0.585	0.575	0.608	0.740	0.196	0.359	0.459	0.532	0.543
Numeral	0.632	1.000	0.800	0.894	1.000	0.778	0.929	0.763	0.892	0.818
Unrecognized word	0.725	0.889	0.032	0.032	0.451	0.305	0.398	0.369	0.472	0.382
Interjection	-	-	0.667	1.000	-	-	0.333	0.667	0.188	0.889

4 Discussion

As it can be seen from the Fig. 1 the best lemmatization results are obtained with the *Semantika.lt* morphological analyzer. Although the difference is very small (i.e., only ∼1.7% points on entire gold-standard corpus), it is still statistically significant. The superiority of *Semantika.lt* over *Lemuoklis* is especially apparent on fiction and periodical texts. The fiction is usually characterized by a high abundance of words, whereas periodical texts are full of neologisms and specific terminology. Thus, a larger number of headwords incorporated into *Semantika.lt* has an obvious advantage over *Lemuoklis*. Surprisingly *Semantika.lt* slightly underperformed *Lemuoklis* on the scientific texts, but the difference is not statistically significant. The terminology used in the scientific texts is not completely settled: some Anglicisms are more popular than their Lithuanian equivalents, some equivalents in Lithuanian sometimes even does not exist, thus are not recorded in the dictionary.

The left diagram in Fig. 2 presents the coarse-grained annotation results. The difference between the results on the entire gold-standard corpus is ∼2.5%: the largest gap is again on the fiction (∼3.5%) and periodicals (∼3.1%), the smallest – on scientific texts (∼0.4%). In the lemmatization task, the *Semantika.lt* morphological analyzer outperforms *Lemuoklis*, but the difference again is not statistically significant. In the right diagram of Fig. 2, which already presents fine-grained morphological categorization results (determined cases, genders, tenses, etc.), the robustness of *Lemuoklis* over *Semantika.lt* on the scientific texts is already statistically significant (the difference is ∼5.3%). However, on the entire gold-standard corpus (the difference is ∼8.1%) and on the other domains, in particular, fiction (∼17.8%), administrative (∼11.0%), periodicals

(∼8.9%), the superiority of *Semantika.lt* is apparent. The lower accuracy in the fine-grained annotation is due to the complicated disambiguation problem and out-of-vocabulary words. The main drawback of both morphological analyzers is due to out-of-the-vocabulary words: i.e., if analyzer cannot recognize the word and indicate its lemma (leaving in original untouched form), it cannot recognize any other morphological information. Thus, the errors in the first lemmatization stage cause errors in the following morphological annotation stages: part-of-speech recognition and afterwards in the morphological categorization.

The detailed error analysis (see Table 2) reveals some major mistakes. The most complicated issue for both analyzers is the auxiliary words (i.e., conjunctions and particles) which can be assigned to the different parts-of-speech without absolutely clear criteria (by the way, some numerals also face this problem). However, *Semantika.lt* analyzer demonstrates significant improvement for the proper nouns compared to *Lemuoklis*. A very specific mistake of *Lemuoklis* is due to the confusion of one letter abbreviations with one letter interjections (e.g., despite interjections in the upper-case at the end of a direct sentence are very rare).

5 Conclusion and Future Work

This comparative research work disclosed strengths/weaknesses of two the most popular and publicly available Lithuanian morphological analyzers: *Lemuoklis* and *Semantika.lt*. Both analyzers were evaluated on 4 domains of the same gold-standard corpus.

The morphological analyzers *Lemuoklis*/*Semantika.lt* achieved ∼96.3%/∼98.0%, ∼92.8%/∼95.3%, ∼78.7%/∼86.8% of accuracy on the lemmatization, part-of-speech tagging, and annotation of the morphological categories, respectively. Despite *Semantika.lt* was superior over *Lemuoklis* on the entire gold-standard corpus and on the administrative, fiction, and periodical texts; *Lemuoklis* yielded equal performance on the scientific texts and even outperformed *Semantika.lt* on the annotation task of the morphological categories.

The experiments with *Lemuoklis* and *Semantika.lt* were carried out on the normative Lithuanian texts. In the future research we are planning to test their robustness on the challenging types of texts: forum posts, Internet comments, tweets, etc.

Acknowledgments. The authors thank the researchers from LLC Fotonija, especially Virginijus Dadurkevičius, for providing information about the *Semantika.lt* morphological analyzer.

References

1. Agarwal, A., Pramila, Singh, S.P., Kumar, A., Darbari, H.: Morphological analyser for Hindi - a rule based implementation. Int. J. Adv. Comput. Res. **4**(1), 19–25 (2014)
2. Akilan, R., Naganathan, E.R.: Morphological analyzer for classical Tamil texts: a rule-based approach. IJISET - Int. J. Innovative Sci. Eng. Technol. **1**(5), 563–568 (2014)
3. Baisa, V., Suchomel, V.: Large corpora for Turkic languages and unsupervised morphological analysis. In: Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC) (2012)
4. Bickel, B., Comrie, B., Haspelmath, M.: Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-Morpheme Glosses (2008)
5. Bögel, T., Butt, M., Hautli, A., Sulger, S.: Developing a finite-state morphological analyzer for Urdu and Hindi. In: The 6th International Workshop on Finite-State Methods and Natural Language Processing (FSMNLP 2007), pp. 86–96 (2007)
6. den Bosch, A.V., Daelemans, W.: Memory-based morphological analysis. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL 1999), pp. 285–292 (1999)
7. Byrd, R.J., Tzoukermann, E.: Adapting an English morphological analyzer for French. In: Proceedings of the 26th Annual Meeting on Association for Computational Linguistics (ACL 1988), pp. 1–6 (1988)
8. Daudaravičius, V., Rimkutė, E., Utka, A.: Morphological annotation of the Lithuanian corpus. In: Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies (ACL 2007), pp. 94–99 (2007)
9. Gelbukh, A., Sidorov, G.: Approach to construction of automatic morphological analysis systems for inflective languages with little effort. In: Gelbukh, A. (ed.) CICLing 2003. LNCS, vol. 2588, pp. 215–220. Springer, Heidelberg (2003). doi:10.1007/3-540-36456-0_21
10. Jęrzejowicz, P., Strychowski, J.: A neural network based morphological analyser of the natural language. In: Proceedings of the International Conference on Intelligent Information Processing and Web Mining (IIPWM 2005), pp. 199–208 (2005)
11. Karp, D., Schabes, Y., Zaidel, M., Egedi, D.: A freely available wide coverage morphological analyzer for English. In: Proceedings of the 14th Conference on Computational Linguistics, vol. 3, pp. 950–955 (1992)
12. Kessikbayeva, G., Cicekli, I.: A rule based morphological analyzer and a morphological disambiguator for Kazakh language. Linguist. Lit. Stud. **4**(1), 96–104 (2016)
13. Khoufi, N., Boudokhane, M.: Statistical-based system for morphological annotation of Arabic texts. In: Recent Advances in Natural Language Processing (RANLP 2013), pp. 100–106 (2013)
14. Koskenniemi, K.: Two-level model for morphological analysis. In: Proceedings of the International Joint Conferences on Artificial Intelligence Organization (IJCAI 1983), pp. 683–685 (1983)
15. Malladi, D.K., Mannem, P.: Statistical morphological analyzer for Hindi. In: International Joint Conference on Natural Language Processing (IJCNLP 2013), pp. 1007–1011 (2013)
16. McNemar, Q.M.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947)

17. Pauw, G.D., de Schryver, G.M.: Improving the computational morphological analysis of a Swahili corpus for lexicographic purposes. Lexikos **18**, 303–318 (2008)
18. Rimkutė, E.: Morfologinio daugiareikšmiškumo ribojimas kompiuteriniame tekstyne [The Limitation of the Morphological Disambiguation in the Digitalized Corpus] (in Lithuanian). Ph.D. thesis, Vytautas Magnus University (2006)
19. Russell, G.J., Pulman, S.G., Ritchie, G.D., Black, A.W.: A dictionary and morphological analyser for English. In: Proceedings of the 11th Conference on Computational Linguistics (COLING 1986), pp. 277–279 (1986)
20. Savickienė, I., Kempe, V., Brooks, P.J.: Acquisition of gender agreement in Lithuanian: exploring the effect of diminutive usage in an elicited production task. J. Child Lang. **36**, 477–494 (2009)
21. Žilinskienė, V.: Lietuvių kalbos dažninis žodynas [The Frequency Dictionary of the Lithuanian Language] (1990). (in Lithuanian)
22. Zinkevičius, V.: Lemuoklis - morfologinei analizei [Morphological analysis with Lemuoklis]. In: Gudaitis, L. (ed.) Darbai ir Dienos, vol. 24, pp. 246–273 (2000) (in Lithuanian)